الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic and Popular Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة مصطفى سطمبولي معسكر
University of Mascara
كلية الآداب و اللغات
Faculty of Letters and Languages
قسم اللغة و الأدب الانجليزي
Department of English Language and Literature

Towards an Effective Assessing Procedure of the EFL Learners' Writing

A Teacher's Guide

Academic Year/ 2022-2023

Preamble

The present enquiry is an attempt to illuminate a significant issue in the learning of the writing skills namely the assessment effectiveness. The spur behind such extravagant wish to accomplish this modest work is certainly the sharp deficits in the writing skills of our EFL learners, in particular the hindrances that impede our learners from enhancing their writing performances. The assumption is that such deficits are due to the ineffectiveness of the writing teacher's feedback. Such feedback ineffectiveness in its turn is unquestionably due to the ineptness of the assessing process of the EFL writing teacher. The study aims at finding a solution to the problem of assessing the written language by providing hypotheses that can link the reliability of assessing with the learners' writing enhancement. The hypotheses call for the design of an analytic assessing system which can control the procedure. Besides, the hypotheses advocate the use of the machine to achieve a better objectivity and reliability of the essay assessment. The proposed hypotheses are attempts to make the assessing of an essay more objective, formal, effective and reliable. Such an assessment may increase the learners' consciousness over writing complications and provide them with positive feedback over the quality of their writing. Undoubtedly, writing is the final process of each well determined unit in teaching. Learners are required to re-invest the thematic and language element acquired throughout this teaching part, by foregrounding a particular function like advising, comparing, informing, etc. Most importantly; writing helps the teacher design an effective feedback. Such a feedback should provide a kind of a channel for the teachers to communicate constructively with students and help them develop as writers. Accordingly, the present work includes two parts. The First part highlights the issue of gauging the written production through subjective test namely the essay, the features of such a subjective test, the holistic approach to the essay assessment, and its alternatives mainly the analytic assessing method. On the other hand, part two portrays and scrutinises the analytic assessing scheme through a Linguistic essay example provided.

Table of Content

Part Two/ Designing an Effective Analytical Assessing Scheme to Gauge Writing

Preface

Writing seems to be the trickiest talent to acquire. You may read, speak, understand a language perfectly, but you may not write correctly. Learning to write is to a great extents dependent upon the effectiveness of the teachers' feedback. In its turn this latter is completely based on the validity and reliability of the kind of assessment provided. The conventional assessment of the learners' writing still has a justifiable lay in the English language skills, but could not be the secure means of assessing writing. Assessment is imperative for an obvious and reliable portrait of how learners are recovering and how fine the processes of instructions deal with the learners' requests. The assessment of the written production can take two shapes either holistic or analytic.

In holistic evaluation, the written production is read for a broad judgement and according to such a judgment the teacher assesses his learner's levels of proficiency. All the facets of the composition (content and conventions) influence the teacher's response, but none of them is particularly recognized or frankly addressed through a checklist. Teachers can never graph students' progress and collect information that will assist them in defining rate and worth. This approach is speedy and proficient in reviewing and checking the overall performance. It may, yet, be unsuitable for estimating how well learners have applied precise standards or developed a particular form.

The analytic assessment, on the other hand, regard writing as to be composed of a range of characteristics, such as relevance, grammar, organization of ideas, expression of concepts, and punctuation, each of which is to be assessed independently. The essay is to be read and measured according to a prearranged list of criteria and a set of principles. The teachers can design reliable assessing system based on selected criteria of evaluation in order to assess each written production aspect alone. Accordingly, the assessment method can guarantee a positive and constructive feedback. Analytic assessing could help the teacher keep the full of writing features in mind as he assesses the written language. It also let the learners

to see areas in their essays that need work when accompanied with written commentary and remedy. Its diagnostic nature offers learners a road map for an effective enhancement and perfection. However, before embarking on the analytic assessment, the writing teacher is to select the assessing components. Many researchers agreed that the ability to write encompasses minimally six ingredients mainly the grammatical skill or the ability to write English in grammatically correct mode, the Lexical talent or the skills to properly select and employ the words, the mechanical cleverness or the aptitude to rightly use punctuation, spelling, capitalisation, etc., the stylistic proficiency or the faculty to use sentences and paragraphs fittingly , the organizational skill or the capability to arrange the written work according to the conventions of English , including the order and range of material, and finally the judgement of appropriateness or the ability to make judgements about what fitting depending on the task, the purpose of the writing, and the audience.

However, the analytic procedure as opposed to the holistic one is recognized as time consuming. The teachers who assess analytically are usually required to make as many separate judgments about one piece of writing. The holistic review is characterised by its easiness. The analytic assessment appears impossible to attain such attribute. The writing teacher will spare a great deal of time and effort because of the limitless criteria of evaluation he has to assess in each essay.

Part One/ Assessing the Written Production, an overview

1. Introduction

Evaluating students' answers to an essay-type question has become a major problem that is still a matter of investigation. The scoring procedure is subjective and completely holistic in the sense that it seems impossible to control the marks that would be given. The teacher's assessment is rather affected by personal opinion and prejudice which may results in scores unreliability. Such a factor is still a problematic in testing and scoring written language. In contrast, objective test such as cloze test items are rather analytic and scientific in a more precise sense. The assessment is always reliable and easy. The teacher's scoring is, at some extent, based on a countable procedure. No judgement is needed on the part of the classroom teacher or other scorer who may be asked to give his opinion on the answer. The marks allotted remain static from a paper to another, and from scoring settings to another. This is due to the fact that the answer needed on the part of the students is either right or wrong.

In any consideration of classroom written tests, a distinction must be drawn between the rather objective tests and subjective ones. Such tests are generally prepared, administered, and assessed by the same teacher. They are mainly used to measure the students' achievements of the instructional goals. Achievement tests, as a fact, are used to indicate group of individual progress toward the instructional objectives of a specific study. They measure the extent to which students have mastered the specific skills or body of information acquired in a formal learning situation. Examples of these tests are final examinations in a course of study. They have a single cut off point: the examinee either 'pass' or 'fail' the test, and the degree of success or failure is deemed important to both the examinees and the examiner.

Objective tests are considered as formal classroom tests because they are analytical, and need no subjective judgements on the part of the examiner. The scoring procedure is highly reliable, and quantified in mathematically precise terms. On the other hand, Subjective tests, such as an essay-type test, are holistic and impressionistic. They have been recognised as informal tests because they are based on an adhoc basis. The assessing procedure is often unreliable. The possibility to achieve scores reliability is still a matter of investigation.

The search for an objective and reliable assessing procedure to the testing of essay-type test has led some examiners to adopt different types of methods. The assessing of an essay-type examination, thus, has witnessed the emergence of four different methods: Short-Type Essays, Multiple-Scoring Method, Error-Count Method and the Analytic Scoring Method. These methods differ from each other in terms of the assessing procedures, but they have a unique common objective namely 'scores reliability' that is extremely impossible to achieve with the single holistic assessing procedure. But the question is still raised: Is it possible to come to a formal and objective assessing practice? Is it possible to achieve assessment reliability with one of these methods?

2. The Study Objective

The principal theme running through this research is effectiveness in assessing students' writing skills. The assumption is that the effective assessment is to be built upon a careful specification of an analytical scoring instrument, which can increase the consistency of the assessing procedure and scrutinise the learners' competences in a chirurgical. Thus, in order to increase the reliability of the scoring procedure it is an indispensable condition to take an over-restrictive view of what it is supposed to be tested (validity). Reliability does not guarantee validity, but it is a necessary prerequisite. As a general rule increased reliability will always increase validity.

Such a research aims at achieving reliability-validity tension Davies (1978), which are the most significant features of the psychometric tests as opposed to those of pre-scientific

days which are highly subjective and unreliable**.** Heaton (1991:16). Its main and fundamental objective is to come to an effective assessing procedure. Such effectiveness is determined by consistency in scoring. Accordingly, the written essay is supposed to be given a trustworthy assessment whenever it is scored and whoever scores it.

3. The Problematic

Scoring an essay test is subjective is the sense that its merit has to be evaluated or judged by the tester who is the test designer and the class teacher at the same time. Both students and the teacher regard the interpretation of the answer and the way it must be tackled from a highly subjective opinion. The students' responses to the subject-matter are entirely subjective in the sense that each of them approaches the essay-question from a different subjective perspective and adopts a highly different strategy. In a sense, each student expresses his own opinion freely using his own words, interprets information in a way he wants and adopts a personal organisation of his ideas.

The teacher, on the other hand, finds himself exposed to with different types of answers. Each answer differs from the others in terms of content, style, and organisation. His judgement of the correctness of the student's response is influenced by his opinion of its content, logical structure or whether he agrees with what the student has written. In such a situation, the judgement of the quality of the essay is extremely difficult. The teacher cannot avoid subjective interpretation of its content or writing process. He finds himself unable to give a conclusive reason to the marks he gives. His decision to the kind of mark to be given to each essay seems inconsistent. The same written production, as a fact, is likely to be given different marks in different occasions. Furthermore, it is probable that different teachers assessing one essay may provide different comments, interpretations and even different marks. Their judgement about the correctness of the essay-answer is highly subjective in the sense that each of them judges the quality of the essay-answer according to his knowledge, and the criteria he relies on in his scoring procedure. The major factors that result from the

scoring of an essay-type test are unreliability in scoring. Also, the students may ignore everything about the how and the why of the grades they may be given. Such factors reflect the real existence of subjectivity in assessing a written production.

As opposed to subjective test such as the essay-question, objectivity and reliability are quite high in the so-called objective tests such as cloze-test items, or true-false answers. The relevance of such kinds of tests is that there is no subjective judgement on the part of the teacher, whose assessing procedure seems very consistent whatever the circumstances are. Moreover, the same test can have the same rank even if it is assessed by different teachers. There is always an objective judgement of the answer. The answer is either right or wrong. No personal interpretations or thoughts are needed from the part of the students, who are supposed to complete the assignment by using the language of the teachers not of their own . Accordingly, the assessing procedure is conducted as objectively as possible. No space is left for subjective judgements.

4. Testing the Written Production

Testing written language can be seen from two different types of tests: objective tests and subjective ones. The main difference between these two tests is mainly recognised in the assessing procedure. Objective tests are always as referred to as standardised tests. The main common testing devices are cloze tests, and multiple-choice items. They had been used as language testing techniques after the psychometric revolution of the 1930s. Standardised tests were developed into writing assessment in order to overcome a number of the weaknesses that covered the testing of essays. The most serious of these problems is assessment unreliability. But evidence has shown that even if reliability is neatly perfect in the case of multiple-choice items where the assessor seems very often like a machine, the essay test is likely to be a very important testing device. From a historical perspective, the widespread use of an essay as a testing device probably grew out of the back to basics movement which emerged in response to charge that many of the educational systems lacked the fundamental academic skills of

writing. The purpose of essay tests was to integrate educational tests more meaningfully into instructional process by emphasising the importance of communicative language testing as a remedy and as a substitute to the psychometric movement which occurred in the period from approximately the mid-1960s to the early 1980s.

The psychometric approach with its standardised tests such as multiple-choice items, and gap-filling and its emphasis on the twin concepts validity and reliability emerged as a reaction against the traditional testing of an essay which was regarded as highly subjective and unreliable. However, by the emergence of the modern approach in language testing, some instructors tried to develop a new technique that should accompany it; (Wilkins; 1979: 82). Clearly, the responsibility for ensuring a better and acceptable assessing procedure depends to a larger extent on the appropriate selection of the measurement device, which may elicit the modern evaluation of the writing skill.

The birth of the psychometric and the integrative approaches, as a fact, has guaranteed two factors: objectivity and reliability. However, evidence shows that these approaches ignored totally the real assessment of the writing skill. Their main purpose is to avoid the side effects which enclosed the measurement of such a type of task. The development of the modern approach, however, has brought a new flavour to writing assessment. The essay test has given a formal status to language testing. It has been thought as being the most appropriate technique that can make student demonstrate their abilities in writing. Such a test emphasises the importance of language performance that is to say, testing the communicative aspects of language such as the content, style, organisation of ideas, and paragraphing, more than language competence which emphasises the testing mastery of language such as grammar, vocabulary and mechanics. It provides the students with an opportunity to demonstrate their abilities to organise language material using their own words and ideas and to communicate. This advantage of an essay would tell a lot about the student's ability in a

particular skill (Nolasco & Arthur; 1983: 17). It also requires free-responses on the part of the learners rather than the selection of correct answers.

The teacher's assessing procedure in mainly dependent upon the elaboration of two indispensable tests: attainment and proficiency test. The former is one which aims to measure how much a learner has learned of what he has been taught, and the latter aims to measure a learner's knowledge of the whole language (Corder; 1985:369). The quality of the essay answer can be regarded from two sides: the answer to the question (the substance of writing), and language form. It may assure two components: performance and competence. The teacher's objectives in constructing an essay-type question are: (1) to measure the learner's achievement of the instructional goals, i.e. to measure his progress toward the instructional objectives of a specific material, and (2) to measure his specific strengths in each component of the essay i.e. to measure his abilities in grammar, mechanics, organisation of the ideas, style, and so on).

In language testing the term subjectivity refers to the holistic or impressionistic evaluation of the learners' products (essay-answers to a single long question). Such an evaluation in mainly recognised as informal. An informal evaluation is the type of evaluation which is broad and global.

5. Features of Subjective Test (The Essay)

Some studies have shown that the evaluation is considered to be subjective whenever it carries the following characteristics. Some of these characteristics such as scores unreliability cause a serious problem in language testing.

5.1. Unreliability in Assessment

Subjective tests refer to the testing of productive skills such as speaking and writing. An essay test is a knowledgeable device used in most common subjective examination. It is a type of test in which the learner is asked to discuss, enumerate, compare, state, evaluate,

analyse, summarise, or criticise involve writing at specified length. Such a test allows the learner to compose his own relatively free and extended written responses to problems set by the teacher. In foreign-language testing these responses may consist of single paragraphs or may be full essays in which the student is rated not only on his use of grammatical structures and lexicon of the target language but also on his coherent ideas and their organisation. Grades for such free-responses tests may also take into account the learner's employment of the graphic convention namely spelling, punctuation, capitalisation, paragraphing and even handwriting. Such a type of test is highly subjective because teachers cannot hide their personal opinions.

According to Corder (1985: 360), our judgements on an essay or a précis are almost inevitably influenced by our opinions of its content or lexical structure, or in an extreme case, whether we agree with what the writer has said. Such judgements have a negative effect on the scoring procedure. It is quite possible that the same written production will be given different scores in different occasions. Reliability in such a situation tends to low as opposed to objective tests. The assessment is impressionistic, broad and difficult to quantify.

5.2. Lack of Validity

The logical result of the test unreliability is the test invalidity. In other words, if the test lacks reliability it will also lack validity. In brief, the test measures something consistently. It cannot measure anything which seems invalid. We want our tests to measure as accurately as possible what they set out to measure. But if for any reason we cannot place our confidence in the results we get, then we can scarcely regard the tests as valid (Corder, 1985:364). While in language testing, everything is subjectively valid; in technical language it is rather unreliable. There are three main factors which show clearly the invalidity of the test. First, the test is regarded as invalid when it contains technical words that seem impossible to understand (face validity is lacking). Second, it is quite possible that the majority of the students may find the test question difficult to respond to, or they may misinterpret it because

the task they are asked to perform is uncommon to them. The teacher may have not prepared them in advance and they do not know how to tackle such type of tasks. The class session may be designed only for theoretical courses.

The methodology of writing an essay may be dismissed and not taken into account by the classroom teacher who is supposed to measure students' writing abilities. In such a situation the test is considered as invalid even if the question seems clear and has a direct relation with the courses they have been taught. Third, the test lacks validity when the teachers cannot know which criteria he has relied on in his scoring procedure; when his judgements seem based on an adhoc basis. This point is the most important one which shows clearly the close relationship between reliability and validity in testing. The test is deemed informal and invalid when it is not followed by an analytical specification of a scoring scheme. Such an analytical scoring scheme can ensure both validity and reliability at the same time. In the absence of such a device, the scoring procedure will be inevitably broad and inexact. As a fact , When we consider that a learner has done a good translation or a good essay we do not know very precisely what quality or qualities we have mastered and we are far from confident that our measure is a valid one (Corder ; 1985:358).

The specification of the analytical assessing scheme should be done in advance, before setting the test, in order to ensure a correct and reliable testing of the supported criteria. The test would seem valid if it deals with the essay as a kind of verbal communication (Yorkey, 1982:235). All the essay components such as content, form, grammar, vocabulary (word-choice), style, and mechanics must be taken into account. It is not to assume that a given score on language form necessarily allows conclusions to be drawn about the learner's language performance. The whole of communicative event was considerably greater than the sum of its linguistic element (Clark, 1973:432). In language testing, the essay is used to reveal the quality of the student's language performance; his ability to communicate ideas as precisely and correctly as possible; and his savoir-faire to resolve the problem that he is exposed to. The

test will lack its validity if the essay is designed just to reveal the learner's competence in one area, grammar for example, but frequently the student's performance and success in accomplishing the task may be masked by errors and a tired marker who fails to make the necessary effort to respond to the writing as a means of communication (Heaton, 1991:149). If tests only focus on grammar, they will not show the teacher how well students can write in English to express meaning. It is a truism to say that in testing an essay we need to know not simply the student's ability in writing correct English, but also how he can communicate his thoughts.

The assessment needs to engage first and foremost with the communicative purpose and overall coherence and organisation of the student's output, not only with localized errors which should be a secondary concern. Innovative and interesting writing is presented as a problem-solving task which challenges, rather than defeats the students (Hamp-Lyons: 1987). The classroom teacher should set forth the objectives of the task that he wants his students to perform. Everything which may cause a handicap towards students' writing abilities should be clearly cleared up. The students should be well-informed of how to tackle the problem imposed and how their responses are to be scored. In such a situation, it is possible to say that the test is legally and formally acceptable.

Unfortunately, this purely a theoretical conception of how validity must be conceived. In subjective test everything is hidebound by personal bias. Therefore, whenever we encounter such deficiencies in language testing, we declare the test as being subjective and invalid. Validity goes hand in hand with objective tests when reliability tends to be high. Invalidity, on the other hand, follows subjective tests when unreliability covers the assessing procedure. It is the classroom teacher who can validate his test or invalidate it. He is the first responsible for the selection of the test materials which can be constructed either subjectively or objectively. The validity of the test is highly dependent on the manner in which the

instrument is employed. Improper administration can invalidate and impair the performance of individual learners and decrease the efficiency of the assessing procedure.

5.3. Complexity of Interpretation

Interpretation is by definition an explanation of something which is not easily understandable. It is a conclusive result which can be determined by time as true or false. Free-responses tests are classified as subjective because their interpretations are highly difficult. It is readily apparent that the students are allowed to express their answers in their own words in a relatively unstructured testing situation. The interpretation regarding the level of ability or correctness of performance on the test may be subjective. Each learner is likely to approach the test and the task it requires from a slightly different subjective perspective, and to adopt slightly different subjective strategies for completing those tasks. According to Bachman (1990:38), these differences among test takers further complicate the tasks of designing tests and interpreting test scores.

Two interrelated factors are supposed to increase subjectively in testing written responses. Such factors may also cause difficulties in interpreting test scores and essay test-responses. According to Bachman, responding to a subjective test is determined by the use of two writing devices: styles and strategies. Every test taker is able to express his own opinion freely and interpret information in any way he wants. In the same way, the tester is able to evaluate the quality of his own opinion and interpretation as well as the organisation and logic of his opinion. Style and strategies are, then, common factors which the teacher finds as real obstacles in his interpretations. Styles are those general characteristics of both intellectual functioning and personality type that especially pertain and differentiate anyone from someone else. Strategies, on the other hand, are specific methods of approaching a problem or task, mode of operation for achieving a particular end, planned designs for controlling and manipulating information. They are contextualized battle plan (Brown, 1987) that might vary from a moment, day or a year to another. People are not machines, even though they are

supposed to have the same devices such as ‘language-processing device’ or ‘language-learning device’ as put forward by Chomsky. The fact is, of course, that our performance in any task is strictly personal. The learners can never approach the stimulus from a unique viewpoint, even if they are familiar with. They cannot evade supporting their own opinions, rather than that of the teacher. The tester, also, cannot evade supporting his own opinions rather than that of the testee. His subjective interpretation of the learners’ responses is not authoritative but seems somehow intuitive. One of the most important characteristics of intuition is its nonverbalizability. Persons are not able to give much verbal explanation of why they have made such particular decision or solution (Brown; 1987:249). Intuition as a human factor involves a certain kind of risk-taking. In subjective tests, assessors must be willing to risk techniques, methods, or assessment that may produce a vague and biased impressionism. That is precisely the difficulty.

5.4. Simplicity of Formulation

The essay-type test is very economical in terms of skills and time that are required to prepare them. It is just too easy for the tester to take pen in hand and to turn out his question in few lines. Indeed, the test objectives can be used directly on course objectives, and test content derived from specific course content. The classroom teacher has just few words to say in relation to what he has already taught to his students. The test question often includes item words such as discuss, comment, or explain. Quite possible, the classroom teacher may use quotations in terms of questions with specific item words instead of his own words. Such a procedure may decrease the time allowed to select the appropriate test question, and relieve the busy classroom teacher from straining his eyes and time in thinking about the question.

5.5. Simplicity of Administration

Open-ended tests such as the essay questions are commonly recognised for their ease of preparation and administration. Such tests are easily administered because of their short

formats. The test directions are usually given verbally in a short time. The teacher sometimes hands out the test question to ensure that all the students receive a better wording of the question. The directions are paraphrased verbally in order to have a better understanding of the task.

5.6. Intricacy of Scoring

In its extreme form, subjective tests are ones where the tester's scoring derives completely from his intuition or personal opinion (holistic scoring procedure). Judgements are rather impressionistic and difficult to quantify. They are rendered in rather global terms (Brown; 1987:249). In such tests, i.e. open-ended tests, the scoring is not easy to achieve since writing proficiency involves numerous traits that seem difficult to define. The assessor makes personal judgements about the quality of the learners' responses. His assessment seems based on fallible opinions, which are mainly affected by extra-factors such as fatigue, carelessness, and prejudice. Such factors influences both what is to be tested and how testing should be carried out. Studies in such a field have shown that the holistic procedure has become the best known one associated with the writing assessment. Its application on the testing ground has not been well embraced by some instructors. Such a holistic scoring is mainly based on an unplanned basis. Although, it seems much faster, it needs a more planned and structured assessing scheme, which can specify in an objective sense test scores. The holistic approach has been adopted to gain a general impression of the learner's responses, but some experiments in the field of testing have shown that such an approach can never achieve reliability. The chief difficulties that have been encountered in using essays as a measurement device are: (1) eliciting the specific criteria that the teacher particularly wishes to test; (2) finding a way to evaluate these free-responses reliably and economically, and (3) making students know exactly why and how they are given such grade.

It is almost a truism to say that each skill is uniquely difficult, but testing essays in itself is a real problem. The teacher should make decision about the matter of control

(objectivity of the evaluation). The holistic method cannot solve the problem of scores unreliability if the teacher must make a judgement about the correctness of the response based on his subjective interpretation of the scoring criteria (Bachman; 199:76). The ability to respond to an essay involves some of the writing skills such grammatical ability, lexical ability, mechanical ability (punctuation, spelling, capitalisation...), stylistic skills, organisational skills judgements of the appropriacy (Kenji. K. & S. Kathleen. K.; 1999). Such skills should be taken into account when responding to students' written production. Perhaps the most difficult and important of these skills is the judgements of the appropriacy.

The assessing procedure needs to be adapted to a logical specification of a scoring scheme. The use of descriptors for each level of the scoring scheme can at least help make the scoring consistent and easy. One possibility is to make an analytical scoring scheme for the overall quality of the writing, but the problem is that, for example, the grammar can be good and easy to score; but the other components poor. It is perhaps more useful to have different sets of descriptors for each aspects of writing that you want to consider. You might want to have descriptors for grammatical correctness, use of vocabulary, content, organisation, and mechanics. These categories might be weighted differently, depending on what you want to emphasise. This is one of the most controlled ways of testing writing. It may direct the teacher's attention to the desired criteria of evaluation; it may also raise the teacher's doubt about his assessing procedure, but it can never solve the difficulties that trouble the testing of essays. Difficulty of scoring is still a problem even if an analytical procedure is applied. The ease of scoring may be achieved if the learner's response will be directed to the limitations set by the assignment. Means that the student will find no difficulties if : **(1)** the answer will be clearly written; **(2)** there will be a clear focus on the topic, and **(3)** if the development and focus will be achieved by means of logically and meaningfully sequences paragraphs. Such requirements, however, appear more imaginative and unlikely to be realised. In reality, the

learners' responses can never meet the teacher's wants, and this is the inevitable problem that the teacher usually faces and which makes his assessment difficult to control.

6. Holistic Approach to Essay Assessment

The assessment of an essay-type test has had different names in language testing. It is frequently referred to as subjective, holistic or impressionistic. Such a method seems to have been established independently in two similar forms in Great Britain and the United States, by Wiseman and his colleagues and known at that time as the 'Demon Method' (Wiseman, 1949)**,** and by 'the Educational Testing' service in the United States, better known through the work of Godshalk, Swineford, & Coffman (1966).

In holistic assessment, the essays are collected from test takers, usually responding to quite general question within a limited time. The assessors make a broad judgement of the quality of the answer in a very short time. The holistic scoring is an approach to the whole writing assessment and not only the scoring. Its main objective is to construct what writing is, and what is important that writers should be able to do with the written language. Ideas are found to be salient trait in most contexts, but this is generally judged in the general rather than the specific. In other words, ideas are checked if they are pertinent, convincing, relevant, and of an adequate quality. The holistic assessment, therefore, focuses only on the most salient criteria for the context, and does not claim to assess every facet of writing competence that may appear in the students' writing.

The Holistic Approach Scoring Type is a subjective procedure in which the teacher makes quick judgements on writing samples and assigns an overall score. One advantage with the holistic evaluation is that it is the quickest method for scoring. The essays are read once. There is one strategy that can be used to evaluate the essay. The teacher normally uses an assessing guide, which outlines the features he should address when scoring essays. The teacher matches features listed on the assessing guide to features on the essays and then

assigns the appropriate corresponding grade. Through this procedure the scorer is assumed to insure scores reliability.

7. Holistic Approach Drawbacks

There are a number of problems with holistic scoring. These problems are very serious and may cause a trouble in the field of language testing. The chief among these is that the holistic scoring is not designed to offer correction, feedback or diagnosis (Charney, 1984). The teacher's assessment of the examiner's work is subjective in the sense that its merit has to be evaluated by the examiner (Pelliner, 1970:19). Such an assessment, as it is always claimed to be, rests upon reasons or principles, but the principles of assessment are truth claim in the absence of conclusive grounds.

Experiences have shown that the assessment made by a single teacher who uses the holistic method is very often unreliable. Marks are awarded on the basis of a teacher's overall impression. Credit is not given for specific categories, and a learner's performance is expressed as a single mark or grade. Furthermore, the students are unable to know exactly why and how such marks are given to them. It is also possible that the teacher cannot justify the mark that he may give to any written test. If the assessment is refractory, the net result will be unreliability in assessing and this is the main factor which differentiates subjective scoring from objective one. How can unreliability in assessment be avoided or at least reduced? Is it possible to find a remedy to such a problematic? Is there any method which can ensure a formal and objective assessment of an essay-type test?

8. Alternatives to Holistic Essay Assessment

The single holistic evaluating procedure carries backwash effects on the ground of assessment. It is, therefore, impossible to achieve reliability if our assessing procedures are based on an adhoc foundation. The possible and unique solution according to some experts to get rid of such a type of test (the single long essay) is the administration of short essays.

Moreover, some studies have suggested the use of short essays combined with the use of three different types of assessing methods namely the Multiple-Scoring Method, the Error-Count Method, and the Analytic Scoring Method. These methods may achieve assessment effectiveness when correctly implemented.

9. The Administration of Short Essays

All the methods that can be applied are presumed to be inadequate to solve the problem of unreliability in scoring. The possibility of achieving consistent and valid scoring by any marking scheme, analytic or otherwise is still a problem. The only way to come to firm and reliable results is to avoid such a type of test. According to (Underhill; 1990: 88), one solution preferred by language test writers is to avoid subjective tests together. Such a solution has been also advocated by (Raatz, 1981) who admitted that both the oral interviews and the compositions are not tests simply because they are not objective.

9.1. The Short Essays Advantages

Some specialists propose to substitute long-type essay questions for the short essays. According to (Pelliner; 1970: 28), it is better to set short essays instead of a single long one. Such a method requires the students to respond to more than two compulsory short essay-questions. Some teachers increase the number of questions in their test to four short essay-questions. The time frame allocated to such a test is the same that is required for the single long essay-question. The assessment of the short essays is not the same as that of the single long essay. If the teacher allocates four short essays to his students, the general mark will be divided into four. Each essay will be given five points. The scoring procedure focuses on two criteria: language accuracy and the content (substance of writing). It is based on three steps. First, the teacher starts by reading the whole essay (that is, each essay) in order to gain a general impression of its content. The answer is either correct or incorrect. No additional information is required. The mark that should be given is from 0 to 5. The second step is

referred to as an error-count method. The mark allocated to the essay can be kept as it is if it is free of error or it can be subtracted if it is covered by errors. The teacher may deduct, for instance, one mark from the general mark given to each essay, for each three errors. In doing so, he may differentiate between students' abilities. The third and final step requires the teacher to count the marks given to the four essays.

9.2. The Short Essays Drawbacks

The objection to the administration of long-type essay questions because they are subjective has been rejected because of the existence of some methods which can be applied in language testing, and also because the so-called remedy (that is , the administration of short essays instead of a single long one) has brought to light some embarrassing side effects . Evidence has shown that the inclusion of short essays as a written testing device may cause two serious problems:

First, the construction of essay questions is a problematic in itself. Subjective tests have been described as being easy in terms of construction and administration. One long question, as a matter of fact, is not difficult to construct. However, the teacher, as a test formulator, may find some serious problems in constructing more than two essay questions. It is possible to pay more attention to the construction of one question, but it seems impossible to give the other essay questions similar weight and level of difficulty. The teacher may not be able to consider two extremely important factors: the time required for testing each essay, and the degree of speediness he wishes to build into his test. Let us assume that a maximum of one hour and half has been scheduled for the test. We should, then, divide ninety minutes into the number of the essay questions we wish to set forth, (let's say four essay questions). Each question is assumed to take twenty-two minutes and five seconds. The problem that may happen is that one of these questions may be answered in one hour or more. In some cases, the students consume the allotted time in order to answer just one question. He may also be perplexed by his failure to decide which question to start with. His anxiety will increase when

he feels that the time is too short. The assignments that he is asked to accomplish will be strictly evaluated with a conclusive single score.

Second, the assessing procedure is also another trouble. The teacher tries to find an adequate way to solve the problem of scores unreliability. His task is mainly directed to what is written by the learners whose answers are responses to different types of questions. The short essay questions are administered to achieve two quite possible results: reliability and ease of the assessment process. These two factors may be achieved if no judgement is required on the part of scores (Bachman; 1990:76). The correctness of the test taker's responses must be agreed on by different scorers, whose scores must be identical (that is, objective scoring). An essay-type test, whether it is short or long, is a verbal communication. It requires the learner to use his own language in order to complete the assignment. It seems probable that the learner will be given the same score if the assessment is done by the same teacher. It is also possible that there is no likelihood of that happening.

In point of fact, if the assessment is done one single scorer (the classroom teacher) who uses a marking scheme, the net result will be scores reliability. However, if the scoring is quite holistic, it is generally possible to have different scores for the same written production. Let us assume that short essay-questions can be scored objectively by the same classroom teacher, who prepares in advance his scoring framework. It is possible, in such a situation, to achieve an easy scoring procedure. The single long essay-question has been rejected because it costs more time and effort than objective tests such as cloze-tests or multiple-choice items. Short essays, on the other hand, differ from long ones just in terms of length, (that is, short essay-questions require the students to summarize their answers). The learners are assigned to respond to short essay-questions in no more than one paragraph for each question. Instructions are given beforehand so that each learner can submit his answers to the test's directions. The problem which is supposed to occur is the difficulty to score all the learners' essays over a limited span of time. If we compose all the essays of one essay test, we will get

a long essay, but with different instructions. Accordingly, the classroom teacher will be faced with two possible factors: the length of each answer, and the differences between the topics. Each essay contains a specific subject-matter which may differ from the others in terms of content, organization, language accuracy, purpose and writing process. Each written test, therefore, combines different topics and different writing processes. The differences and the length of the answers are real obstacles that the classroom teacher may encounter and which can really increase rather than reduce the difficulty of scoring.

As a matter of fact, the solution which advocates the use of short essays instead of a long one is pedagogically unacceptable. It is not possible to substitute a problem with another problem. The single long essay-question is presumed to be subjective and difficult to score, but evidence has shown that the inclusion of short essays has entirely proved its handicap in language testing. A good test must be free of any problem that may hinder the testing procedure. The attitude towards the avoidance of the administration of single long essay question because it is subjective has no sense. There seems little reason to exclude single long essay as a testing device simply because it is not objective. The possibility to resolve the subjectivity in scoring an essay-type test cannot be left out. It is possible according to (Nitko, 1983) , (Cronbach, 1984) , & (Gronlund, 1985) to achieve an objective scoring and to reduce scores unreliability, not by excluding the single long essay question , but through using the appropriate method. In language testing, subjective tests such as open-ended tests permit the use of techniques that are natural and seem outwardly very valid.

10. Multiple Scoring Method

The development of Multiple Scoring Method has been motivated by the desire to find ways of assessing writing with the levels of objectivity and to avoid unreliability in scoring. Such an attempt provides some diagnostic information to students and to their teachers that the holistic scoring can never be achieved objectively and more reliably if it is done by one single scorer. The teacher's assessment is regarded as extremely subjective and unreliable. It

is probable that each written test can be given different scores in different occasions. Such unreliability in scores is due to some physical and psychological factors. Such factors are clearly put forward by Heaton: the examiner's work is a highly subjective one based on fallible judgements, affected by fatigue, carelessness, prejudice, etc... (Heaton; 1975: 135)

However, if the assessment is based on fallible judgement, the net result is scores unreliability. Thus, the only way to ensure an objective and reliable judgement of the essay is to enlist services of some equally competent instructors. In foreign language testing the definition of scorers, readers, raters, markers, or judges reflects its use for people who correct other types of language tests; they mark the papers but they never meet the individuals who write them. In a sense, the scorer is a teacher whose task is to correct the written productions of other students and not his own. Such a method may encourage a better scoring procedure.

Multiple scoring method is one aspect of the holistic scoring. The theoretical foundation, upon which the multiple scoring procedure is derived from, is the holistic scoring. Scorers make judgements of the answers as a whole: that they are unable to separate out facets of the essay and identify them. Recently, adaptations have arisen, most notably the developments of essay scales, and or rating guides to accompany the multiple scoring sessions resulting in what is known as 'Modified Holistic Scoring' or 'Focussed Holistic Scoring' , but the holistic scoring still yields only one score to express the quality of the students' essays.

The multiple scoring method implies giving separate scores for one written production in order to obtain the suitable score. It has been thought that it is possible that more than one reader can in fact increase reliability than the single score of a single assessment. According to Ingram, the only way to increase reliability of the marking is to have several judges, whose marks are average (1970: 96). She noted that the judging could be perfectly adequate provided that three examiners judge each essay separately in one occasion. The teachers are asked to read anonymous written productions. In its classic form, the multiple scoring method consists of two readers (raters) scoring the same script, but if the ratings of two readers do not agree,

the paper should be read a third time, and then to accept which ever rating is nearer to the third reader. Such a procedure is used mainly in competitive examinations.

Before each scoring session, scorers are provided with the model essays and assign a rating based on that comparison. The model essays represent borderline cases. Each essay to be rated must, by definition, fall above or below a model. One model essay represents each dividing line. The teachers should read each essay to gain a general impression of its quality in relation to the model essay. They should first make decision of the final score to be given to each essay model. Such a decision may limit ratings in some agreed marks so that no scorer can give whatever he wants to give. It is essential that all scorers be thoroughly familiar with the rating criteria in order to carry on a common scoring procedure. There are four recognisable steps:

Step 1: First reading:

The written production is read quickly by each person in the scoring group until all essays are being read.

Step 2: Initial Scoring:

Scorers then pair up and work together to score an essay based on a scoring guide on a 6 (excellent) to 1 (poor) scale. Initially, the pair determines if a paper's answer is on the topic or not. Papers that do not fit the '1' or '6' categorization are separated from the two extremes. These opposing scorings of '1' and '6' indicate the worst and the best scores.

Step 3: More Specific Scoring:

The next step involves sorting out the essays left out of the '1' or '6' scoring. The readers quickly read this pile and sort the best of this middle pile into the category of '4' and the worst of the pile into the '2' category. Those remaining essays in the middle are the '3s'.

Step 4: Final Scoring:

The scorers now have five distinct piles with an assigned number from ‘1’ to ‘5’ representing worst to the best as described by the traits relevance, paragraphing, style, organization, diction, mechanics, and grammar.

10.1. The Multiple Scoring Method Advantages

The Multiple Scoring Method may, first, achieve objectivity. Objectivity here refers to the maximal increase of different judgements in to maintain a solid and adequate score. One single judgement may affect the score, but more than one can be more objective. The Multiple Scoring Method is an approach to the whole writing assessment and scoring. Such a trend seeks to avoid scores unreliability and to develop an objective scoring procedure. Evidence shows that when the scores on the multiple scoring method are combined to create a single composite score in use in making an administration decision, that single score is highly objective. The use of composite scores can increase objectivity as follow:

Assume that each essay is scored by two scorers. The result is two scores, one matched pair. We may then obtain a single by dividing the pair into two. Because two judges are used, the score will, in fact, be more objective, because it is a combination of two different judgements. Most programs also use a third scorer in cases when the first two scorers are far apart in their judgements; the way these third scores are used vary, but their result is an adjudicated score that is theoretically closer to a true score than the first two scores alone. Multiple scoring method possesses psychometric properties that enhance the objectivity of a score which can be used making yes/no decisions such as whether or not to accept the candidate into a program of study where writing competence is required and for setting cut off points such as the level below which a student should be placed into a remedial writing programme.

Moreover, Multiple Scoring Method may enhance feedback. A key statistical question that must be resolved when using a Multiple Scoring Method is: whether scores should be combined and how. If diagnostic information is part of the purpose of assessment, clearly

each of the scores should be reported separately. If objectivity is a key, multiple scores when combined result in a highly objective scores. The multiple scoring method shows remarkable information of the different sores to be given to one single essay. Scores exist not simply to assign decisions, but also to communicate decisions (Hamp- Lyons, 1992). Scores are information which can be shared with the students, their teachers, and other concerned parties and used by them to take various kinds of action in the context of the information. In contrast with the single holistic scoring where the scorer who notices an unevenness of quality in the writing has no way to report this observation, and must somehow reconcile it as a single score, multiple scoring permits judgements and differences of writing be assessed and reported .

10.2. Multiple Scoring Method Drawbacks

Multiple Scoring Method has also some shortcomings that should be taken into consideration. The Choice of Assessors is a real trouble**.** The classroom testing is always undertaken by one teacher who is in charge of the course session, the test construction and the scoring procedure. These threefold task is restricted to all teachers. The courses are in some cases agreed on by the administration. The testing procedure, on the other hand, is a teacher-control. Each teacher seems authoritarian in his testing. Such an authority is limited to his choice of the question-type and his method in scoring. Accordingly, the possibility to obtain help of other colleagues working on the same ground (sharing sufficient knowledge on the subject matter) remains very low in all cases. It seems unlikely to happen that one teacher can accept to re-correct the students' written productions of other teacher. This does not mean that teachers refuse to co-operate, but there are some reasons that may justify such behaviour. It is a truism to say that each teacher is allowed to correct just his students' scripts. He knows which answer is acceptable and which is not, and how exactly to deal with them. In addition,

the correction may take a great deal of time and great efforts, because of the great number of scripts teachers usually assess.

It is true that the correction of essays is very often difficult. Such a difficulty can be avoided if it is an objective test, such as multiple-choice items, where the answer is either right or wrong. In an essay, each teacher may regard the correctness of the test taker's answer from a slightly personal way. This may cause a disagreement among scorers over the marks they may give.

Besides, there are three main causes of divergences among assessors who independently mark the same set of written productions: (**1)** the marks may differ in average standard or level. One teacher may be generally severe, another completely lenient; (**2)** the marks may differ in their scatter or spread. One teacher may employ the whole range of available scale, another only a part of it; **(3**) the marks may order the learners differently. According to Pelliner; (1970:27), discrepancies among the markers in rank order are reflected in low Inter-correlation among the arrays of marks they assign.

The disagreement has a long historical background. Evidence has shown that one written production or an essay can be given different grades if it is assessed by different teachers. Such a 'number paradox', as it is referred to by Underhill (1990) can never achieve reliability on the ground that the improvement does not represent greater agreement on the value of the essay. People are inconsistent; they do not always agree, either with each other or with what they said or thought last time (Underhill; 1990:89). The number of markers that are supposed to avoid subjectivity and to reach a reliable and objective scoring, can in fact, increase marks unreliability. This '*hiccough'* (inconsistency) is the result of the markers' personal judgements of the essay content, style, organization and procedure. Each teacher may approach the quality of the essay from a highly subjective way. He may regard the essay from a different point of view. He may rely on his own knowledge of the subject-matter, and how it can be responded to.

11. Error-Count Method

The Error-Count Method is not new. It has had different names such as the 'Mechanical Accuracy' or 'Traditional Scoring'. Earlier in this century this method was used for the purpose of achieving reliability in scoring. It has been introduced as an objective method for the scoring of an essay-type test. The teacher's objectives in administering such a type of assessment are, first; to test the students' mastery of the courses taught, and second; to see whether or not they are able to write accurate sentences within an essay-type question paying more attention to the language form.

The procedure in scoring an essay normally goes through two steps. In the first step, the teacher proceeds by reading the whole essay in order to evaluate the quality of the answer. A general mark is, therefore, given to the proposed answer. The step, on the other hand, recalls for the subtraction of marks of, hence 'error-*hunting'.* The procedure, in such a situation, consists of counting the errors made by each examinee and deducing the number from a given total. For example, a student may lose up to 10 marks for grammatical errors, 5 marks for misuse of words, 2 marks for punctuation, etc. The teacher in reducing the number of errors from a given total makes a distinction between two conventional types of errors: major errors and minor ones.

The major errors, or global errors, are those errors which involve the overall structure of a sentence and result in misunderstanding or even failure to understand the message which is being conveyed.

The minor errors, on the other hand, are those errors which cause only minor trouble and confusion in a particular clause or sentence without hindering the reader's comprehension. They are mainly recognized in the misuse of articles, omission of prepositional, lack of agreement between subject and verb or incorrect position of adverb.

Such a distinction seems crucial in the sense it makes balance between errors. The deduction normally follows the value of errors. According to Underhill (1990: 102),

normally, one mark is deducted for each definite error from a starting point of , for example , ten, but sometimes a distinction is made between major error (1 mark off) a minor error (½ mark off).

11.1. The Error-Count Method Advantages

The Error-Count Method has been adopted to determine the effectiveness of the writing skill. The examiner limits his/her focus only on two components: The relevance (the ideas of expression) and language form. The first component is used just as‘bait’ by which the teacher can, on one hand, evaluate the student’s achievement of the courses taught, and on the other hand, to detect the types of errors that learner may make. The second component is given more attention. The accuracy of writing seems to take a major part, more than the relevance. The examinee in not to be penalized for his misinterpretation or his failure to answer the question, but he will be judged for each error he will make even if he is ‘*on’* or ‘*off’* the topic. Such a judgement gives no interest to the purpose of the written test. The examiner’s attention is more influenced by his/ her deliberate hate of errors. Everything that goes beyond such a purpose is to be considered secondary.

11.2. Error-Count Method Drawbacks

To measure content and construct validity test developers must pay more attention to the evidence for what is valued in writing in the context to which the writing test applies, design prompts to elicit that kind of writing and scoring procedures to judge those values and ensures that scores keep values in mind. These judgements of prompts and scoring procedures are in large part content validity judgements. Cronback (1949:48) called this ‘*logical validity’*. This must be compelled with a clear sense of what is involved in the construction of written discourse, of the limitations imposed by the assessment-medium keeping in mind what it means to write in these circumstances.

The Error-Count Method has been developed to satisfy just one objective. The focus on language accuracy is, therefore, the main reason of the assessment. The other aspects of the essay writing are not taken into consideration. The Error-Count Method has proved its failure to achieve a formal assessment of an essay-type test. It has ignored the real purpose of essay writing communication. Construct validity, in such a situation, is completely dismissed. According to Davies (1968), the validity of the test should reflect the principles of a valid theory of foreign language testing.

The essay as commonly conceived is a band of frequencies used for sending out a particular message. These frequencies are mainly organized in a way that let the message goes across. In language testing, such frequencies are recognized as the organization of ideas, style, grammar, mechanics, handwriting, paragraphing, and so on. The message is the content, the substance of writing. If the focus is just on one of them, the assessment will lose its legitimacy. It will be considered as informal and invalid. The Error-Count Method lacks validity in two main points: the focus on language accuracy and the negative treatment of errors. The learners are assigned to respond to an essay-question. Such a type of task, as a fact, needs more emphasis on the whole components of the essay rather than language accuracy.

The students are supposed to write an essay in a way that makes the writing process covered with definite or indefinite errors. They are asked to respond to free-stimulus and to use their own language in order to complete the task. In such a situation, the assessment will be valid if it deals with the whole rather than with the parts. The Error-Count Method is based on the assessment on language accuracy such as grammar, vocabulary and mechanics. Such an approach is still used today by some teachers who favour to concentrate on the negative aspects of the writing task placing the learner in a position that he cannot write for fear of making mistakes**.** According to Ur. P. (1996:171), this over-emphasis on language errors can distract both learner's and teacher's attention from the equally important aspects of content

and organization. The essay test should not be used to assess only specific components such as the mastery of language.

Such an assessment is formally inadequate and informal. The assessing procedure should not be just quantitative (counting errors). Quantitative scoring procedure is unavoidably impractical and some form of qualitative scoring must be formed. In language use the whole is more important than the parts. No matter whether the parts are isolated in terms of structure, lexis or function, it is implausible to derive hard data about actual language performance from test of control of these parts alone. According to Clark (1983:432), the whole of communication event was considerably greater than the sum of its linguistic elements.

The essay-type test should be designed to reveal not simply the language competence, but to reveal the quality of the candidate's language performance. It is not safe to assume that a given score on the former automatically allows conclusions to be drawn about the latter. Frequently, the student's performance and success in accomplishing the task may be masked by errors and a tired marker fail to make the necessary effort to respond to the writing as a means of communication. *The assessment*, according to Harris (1993:121), needs to engage first and foremost with the communicative purpose and overall coherence and organization of the text, not only with localized errors which should be a secondary concern and always related to other primary matters. The teacher or the scorer should be more concerned with the *'positive scoring'* Doff (1995) which gives more emphasis to the content and meaning the student is trying to express. He should judge the learners for what they are able to do rather punish them for what they cannot. Writing, according to Lyons & Heasley (1987), is presented as a problem-solving task which challenges rather than defeats the students. It is that aspect of communication that should be given more importance. The students should not be placed in a situation where they are asked to pay more attention to language competence rather than to language performance. They should be motivated to use their own words, style

and writing strategies. The scoring procedure, on the other hand, should be adopted to the communicative language testing. Obviously, this will have a better effect on the students attitudes to learning.

12. The Analytical Scoring Method

In as much as the preceding methods have been recognized in, somehow , as invalid and informal, researchers advocate a more available and a valid method to be used by a single teacher, who is supposed to be three persons in one: classroom teacher , test designer and test scorer. He is supposed (1) to prepare the course and to deliver it to the whole class; (2) he constructs the question according to what have been already taught, and (3) he evaluates the quality of the test answer with a conclusive mark, according to his knowledge of the subject-matter and in a way that permits the students to know exactly why and how they are given such a mark. In such a situation, there is no need to enlist services of other teachers who may affect the testing procedure, and since most teachers have little opportunity to enlist services of two or three colleagues in marking class composition, the analytic method is recommended for such a purpose Heaton (1975: 137). This method depends on a carefully specification of a marking scheme, which has been carefully drawn by the teacher. Thus, one way of making subjective, impressionistic judgements more objective is to devise a marking scheme through bands and scales in which the judging criteria is described as precisely as possible. These bands should be made as simple as possible (range of vocabulary, grammar, style, appropriateness, etc.) so that the assessor will not have to take into account too many aspects at the same time.

Studies such as those of Bachman (1990) strongly suggest that the assessments made by a single scorer who uses a framework (an analytic scoring scheme) of this sort are more reliable than the global impression assessments of one person. According to him; the test such as the essay that involves the use of rating scales are necessarily objectively scored, since there is feasible way to ‘objectify’ the scoring procedure Bachman (1990: 76).

In fact, evidence has shown that when the scoring procedure is done by only one scorer and when the standard (the analytic scoring device) remains reasonably consistent from an essay to another, the net result is scores reliability.

12.1. The Nature of the Analytic Scoring Method

The Analytic scoring Method comes as a reaction to the Impressionistic (Holistic / Subjective) scoring. It is a psychometric method (Underhill, 1990) which is used to improve the reliability of the scoring of an essay-type test. It depends on the Atomistic Approach (Lado, 1961), which is the breaking down of the complexities of language into isolated segments. According to Morrow (1979: 145), this influenced both what is to be tested and how this testing should be carried out. The teachers, who advocate the use of such a method, support the arguments of the psychometric view that it is possible to reach reliability. The scoring procedure can achieve objectivity although it is subjective in nature.

12.2. The Analytic Scoring Process

The process consists of separating the whole writing process into categories and to mark each category separately. The separate marks are, then, combined to give an overall mark to whole essay-answer. Such a *'counting procedure'* (Pollit. A, 1990) permits the assessor to limit his assessment to the 'marking *protocols'* (Underhill, 1990) that he has already selected and graded before the administration of the test. Accordingly, *each student is* able to see how his particular grade has been obtained (Heaton; 1975: 137).

This is, indeed, a crucial point which differentiates the Analytic Scoring Procedure with the Holistic one. The teacher can easily convince his students about their performance. It is also probable to achieve agreements among different assessors if equipped with the same scheme and with the same procedure.

12.3. The Analytic Scoring Method Drawbacks

The Analytic Scoring Method is also counterbalanced by some hindrances. There are, however, some frequent problems that may not affect the assessing procedure but rather the assessor himself. Studies such as those of Ingram (1970), Madsen (1983) and Peter. W. Foltz, Darrel. Laham & Thomas. K . Landauer (2000) protested against the Analytic Scoring Method. They claim that the scoring procedure made by a single scorer who uses an analytic framework is entirely adhoc and difficult to score. They, therefore, noted two existing problems: (1) a problem with the specification of the scoring scheme, and a (2) problem with the scoring time frame.

As far as the specification of the scoring scheme is concerned, some teachers find that such an analytic device is quite informal. It seems to them that the Impressionistic is better and formal in use. According to Ingram (1970:96), there is no evidence that this is any more valid and reliable than the overall impression marking of experiences examiners. Such a claim is supported by Madsen, (1983) who strongly criticized the Wiseman method of achieving reliability on the ground that the analytic approaches to the scoring of an essay are not well-specified and do not represent greater agreements of how to weight each area of the essay. According to him, a major problem with the analytic approaches is that one never knows exactly how to weight each error, or even each area being penalized. (Madsen; 1983: 21)

The Analytic Method has been claimed to satisfy two interrelated approaches. The first approach is mainly recognized as the Points-Off Approach. It is based on two steps. The scorer starts first with a grade and then reduces it. In the second step the scorer gives points for acceptable work. If we combine the two approaches we obtain the following step-by-step procedure:

(1) The scorer divides the general mark (20/20) into the number of components he wants to evaluate.

(2) The scorer reads the whole essay-answer in order to gain a general impression of its content. The answer is considered either right or wrong according to the teacher's previous knowledge of the subject- matter. Some essay-answers will be accepted because they are on the subject, and some others are to be rejected because the respondents are off-topic.

(3) The scorer moves to the second step when the answer in on the subject. He tries then to evaluate the quality of the other components that he has already selected. Each of them is allotted a specific mark.

(4) The scorer counts errors made in language accuracy (grammar, mechanics, vocabulary...).

(5) The scorer deduces the number of errors made by the examinee from the specific mark allotted to each area (component).

(6) The scorer counts all the marks that he gives to all the selected components, in order to give a general mark to the whole answer.

Some examiners regard such a procedure as being invalid and time consuming. The scorer will spend more time in scoring. Such a factor, according to them, may be avoided in using the Impressionistic (Holistic) Method which is less time consuming. The scorer will be able to correct as much exam-papers as possible.

12.4. The Formal Status of the Analytic Scoring Method

The Analytic Scoring Method, even if it is claimed to be difficult to apply, it is the most useful method that can achieve a very formal scoring. The pessimistic view raised against it has no sense in language testing. According to Bachman & Palmer (1996), the 'usefulness' of a classroom test can be determined by considering the test's reliability, security, and feedback. These characteristics form the main basis of a good test.

The Analytic Scoring Method has been advocated in order to achieve objectivity in assessing the learners' written production. Experiences, for such a fact, have shown that one way to improve the reliability of the Impressionistic (holistic/ subjective) scoring is to adopt an atomistic/analytic procedure. Such a method brings to light some considerable results that

can never be realized with the other mechanical and holistic methods. The findings resulting from the application of the Analytic procedure indicate that it is possible to achieve the following advantages:

(1) Scores Reliability: Scores change from one person to another when there is no standard procedure to rely on. Reliability can be increased if we make use of an analytical scoring scheme. Such a device can firmly maintain the results. According to Pelliner (1970: 28), studies such as those of CAST (1939) strongly suggest that the assessments made by a single marker who uses a framework of this sort are more reliable than that global impression assessment made by one person. It is possible, accordingly, to have more than one scorer giving the same score to the same written production in different occasions. Such a conclusive assessment is due to the fact that the test assessor takes an over-restrictive view of what it is that he is testing. The scoring procedure, in such a situation yields data which is easily quantifiable.

(2) Security: The only way to make the learners feel at ease is to make them know how they are going to be assessed and scored. The specification of a consistent scoring scheme can easily make the classroom teacher control his assessing process. At the same time, each student is able to see how his particular grade has been obtained (Heaton; 1975: 137).

(3) Feedback: The students know pretty how they have obtained such grades. Accordingly, The scoring system gives them meaningful feedback on various aspects of their performance. They will be encouraged to prepare language for the test and to pay more attention to how to respond to the question. The experience tells us that the students view the test as meaningful useful, meaningful, and fair if they feel that the scoring which accompanies it encourages using their own words rather than punishing them for something they do not know.

13. Conclusion

The quest for an objective scoring procedure which can achieve a conclusive and a reliable score is still a problem that should be resolved. The essay-type test, even if it is a very

valid and efficient testing device, has contributed greatly to our understanding of the effects of subjective judgements on the scoring procedure. Subjective written test such as the essay has brought to light a very serious problem that impels testers to adopt different strategies in order to avoid it. Such a problem, known in language testing as scores unreliability, is one of the main results of the single holistic scoring or more precisely the subjective scoring of a written discourse. While there is nothing that anyone can do to avoid such a problematic, there are some methods that can be applied to reduce the effects of subjective judgements.

The inclusion of the four types of the scoring methods, such as the Essays-Questions type Method, the Multiple-Holistic Scoring Method, the Error-Count Method, and the Analytic Scoring Method, into the field of testing written language, such as the essay-type test, have really clarified the problem of subjectivity in scoring an essay-type examination. Such methods work around a common point - avoiding or reducing subjective judgements in the hope to achieve a more exact and reliable scoring procedure, yet objectivity and reliability are still very far to reach.

The possibility to come to an objective scoring as it is the case with objective tests can be achieved just by the Analytic Method, which is quite quantified in a mathematically precise sense. Such a method is assumed to be the only formal type of methods which may maintain validity in relation with reliability. The other methods, on the other hand, have not been satisfactory enough. Some of them have lacked validity, even if they have come in some ways to conclusive results. Some others have lacked reliability, even if they have avoided subjectivity. Objectivity in scoring a written production can be maintained if there is a high probability to bridge the gap between both reliability and validity. Such an attempt is likely to be realised if we make an adequate use of the Analytic Scoring Method, which is claimed to be as the more appropriate testing method that may reach a more conclusive and more formal scoring procedure.

Part Two/ Designing an Effective Analytical Assessing Scheme to Gauge Writing

1. Introduction

It is often conventionally assumed that tests are mostly used for assessment: the test gives a score which is assumed to define the level of knowledge of the testee. This may be in order to decide whether he can be placed in appropriate language classes. Such tests or rather the scores given in any tests can give the teacher information about his students' achievement of the instructional goals. They can also give the degree of success not of individuals but the instructional program itself. Teachers are assumed to prepare the test in the same way they do with the courses. There are now a number of very excellent textbooks on the methods of teaching English as a second or foreign language. There also many testing methods which are mostly used for objective tests, but the field of testing is still lacking a short, concise text on the testing of subjective test such as the essay, a subject which is still a problematic.

The analytic scoring method has proved to be the most objective method that can to some extent reach a conclusive and reliable score. This may be because it is based on a psychometric /atomistic approach. Such a method can arrive at suitable results if they are effectively used. A formal scoring procedure needs an increasing concern from the part of the classroom teacher who is assumed to be in charge of the construction and the administration of the test question. If the construction of an essay question needs some steps such as planning the test, preparing the test items and directions; if the administration of such a question includes fairly detailed instructions in order to validate the test process, the scoring procedure must also be based on an organised techniques that should carried out in order to achieve reliable scores. The reliability of the test scores is highly dependent upon the manner in which the scoring procedure is employed.

The scoring procedure, accordingly, is based on four considerable steps: the selection of the scoring criteria, the specification of the criteria of evaluation, the specification of a

scoring scheme and the selection of the test question. Such steps are organised in a way that permit the classroom teacher to be more objective in his judgements of the students' responses.

2. The Formal Scoring Procedure

Before any attempt to respond to students' answers there are steps that the teacher has to take into account: (1) the selection of the scoring components, and (2) the criteria for grading.

3. The Selection of the Scoring Components

An essay is a type of test in which the learner is asked to respond to the essay question set by the teacher using his own words. The response should consist of all the essay components. The learners are to be rated not only on the direct answer (relevance), but also on their use of all the components which combine an essay. The relevance of the key items such as discuss, comment, explain, compare, analyse, enumerate, evaluate, state, etc are used to direct the students' attention to the type of essay they should deal with.

The ability to write an essay involves many interrelated components. These components lead to clear, fluent, and effective communication of ideas. Most instructors and experts in the testing of English as a foreign language would probably agree in recognising a number of diverse elements which constitute the fundamental pillars of the writing process. Such elements may differ from person to person, but they all agree on the most important of them. The other elements, however, are considered secondary.

To start with, Lado sees that there are things that can measured in connection with content (1) the points of information to brought out ; (2) the organisation and sequence in which these points are presented ; (3) the formal signals given the reader to guide him in understanding the topic fully (Lado. R; 1962: 248). The writing process as it is defined by Lado includes three major components: (1) relevance (2) organisation of ideas and (3)

paragraphing. These components are vital but there are also two other indispensable ones: Language accuracy and style .We will make more progress, according to him, by measuring language as language , and content and style as content and style. What is, therefore, vitally important in an essay are these five major components: relevance, organisation of ideas, paragraphing, language and style.

Besides, language and style are clearly defined by Harris. D.P. (1969: 68-69) who believes that the writing process includes five general components:

- Content: The substance of the writing: the ideas expressed
- Form: The organisation of the content
- Grammar: The employment of the grammatical forms and syntactic patterns
- Style: The choice of structures and lexical items to give a particular tone or flavour to the writing
- Mechanics: The use of the graphic conventions of the language: spelling, punctuation, and capitalization

From the above we can recognise that Harris adopts the same writing process that has been introduced by Lado. Special attention has been given to language and style. These two components are clearly identified by Harris. Language, for example, includes both grammar and mechanics, and style includes diction and writing fluency. The form as it is introduced by Harris is clearly identified by Lado as the organisation of ideas into paragraphs. According to Lado (1962) and Harris (1969) the writing process includes: (1) the content (the substance of writing); (2) organisation of ideas; (3) paragraphing; (4) grammar; (5) mechanics and (6) style. These six components have been probably recognised as the most important elements of an essay.

Moreover, Raimes (1981) sees that the writing process combines many interrelated components. He categorises them as content, organisation, grammar, syntax, mechanics, word

choice, purpose, audience, and the writer's process. According to him, all these components lead to clear, fluent, and effective communication of ideas.

As opposed to Lado (1962) and Harris (1969) , Raimes (1981) adds three possible components : (1) the purpose of writing, (2) the audience and (3) the writer's process. Such components are mainly used in writing a free-composition. A composition test allows the student to compose his own relatively free and extended written responses to problems set by the teacher, (Harris; 1969: 05). The essay test, on the other hand, requires the learner to respond to the question-matter in an essay framework. The two different tests share the same six components cited above. The main difference is made when the purpose of the test is set forth.

In addition, Yorkey (1982) gives special emphasis to one particular aspects of an essay test. According to him; an essay-type examination is a verbal communication. The clarity of the message depends upon the clarity of your expression. If your grammar is imprecise, if your vocabulary is ambiguous, if your organization is distorted, if your handwriting is illegible, there is likely to be a breakdown in communication. Even if the message Comes through confused but comprehensible, your teacher may unconsciously deduct for straining their eyesight and patience (Yorkey; 1982: 236). According to Yorkey the writing process should include six components: (1) the message, (2) the style (written expression), (3) grammar (4) vocabulary, (5) organisation, and (6) handwriting .The main difference between all the writing processes recognised by Lado (1962), Harris (1969), Raimes (1981) and those that Yorkey (1982) introduces in the addition of the element of handwriting. Yorkey regards handwriting as an integral part of the writing process. The clarity of the message, according to him, is mainly dependent upon the clarity of the handwriting. It is, therefore, an, indispensable component that should be given a special emphasis.

Furthermore, and according to Madsen (1983) there are eight components of the effective written production:

- Mechanics : (Spelling , punctuation , and capitalization)
- Vocabulary
- Grammar
- Appropriate content
- Diction (word selection)
- Rhetorical matter (organisation , cohesion , unity)
- Logic
- Style

In their turn Dickens and Germaine (1993: 51) recognise twelve components that can be included in any writing process. Such components are of two types. The first type includes those that focus on the accuracy of language use, such as grammar, vocabulary and tenses. The second includes those that focus on communication such as style, appropriateness, effort to communicate, fluency, and relevance of the content.

Also, Penny Ur (1996:163) admits that the purpose of writing according to is the expression of ideas, the conveying of a message to the reader, so the ideas themselves should arguably be seen as the most aspect of the writing. On the other hand, the writer needs also to pay more attention to formal aspects: neat handwriting, correct spelling and punctuation, as well as acceptable grammar and careful selection of vocabulary. In such a situation, the essay includes two interrelated aspects: language form and content. The former includes spelling, grammar, punctuation and handwriting. The latter, on the other hand, includes interest, originality of ideas, effectiveness of expression and organisation.

Moreover, Kenji. K. & S. Kathleen K. (1999) think that the writing process, as commonly conceived, is a highly sophisticated skill combining a number of diverse elements, only some of which are strictly linguistic. The ability to write involves at least six components. They are:

- Grammatical Ability: This is the ability to write English in grammatically correct sentences.
- Lexical Ability: The ability to choose words that are correct and used appropriately.
- Mechanical Ability: The ability to correctly use punctuation, spelling, capitalisation, etc.
- Stylistic Skills: The ability to use sentences and paragraphs appropriately.
- Organizational Skills: The ability to organise written work according to the conventions of English, including the order and selection of material.
- Judgements of Appropriacy: The ability to make judgements about what appropriate depending on the task, the purpose of the writing, and the audience. These two latter are mainly recognised in a composition test.

The foregoing surveys show clearly the complexity of the writing skill. Such a task, as commonly conceived, is a well-developed device which would require more attention and a careful specification of its components. Obviously, these components constitute the primary criteria of evaluation that the teacher would take into account in order to keep a consistent testing procedure. It would be of a paramount importance to select beforehand what is to be tested in an essay. According to Dickens & Germaine, Teachers should choose in advance of administering a test which criteria they will use to mark learner's work." (Dickens & Germaine; 1993: 50)

A careful selection, therefore, seem to lead to a better understanding of the main purpose of an essay test. The essay-type test is a kind of verbal communication which needs all the properties which constitute its fundamental basis. The selection of the components of

evaluation, as it is advocated by Mahili (1994), is the most crucial fact that the assessor has to take into account before starting the scoring procedure. According to him, some teachers tend to impose themselves as authorities and make comments reflecting the application of an ideal standard rather than having a set of criteria for marking." (Mahili. I.; 1994: 24)

Accordingly, the criteria of evaluation have great effects on scores. It has been found that among the drawbacks of the holistic scoring method is the lack, or the unreliability of the criteria of evaluation. The indecision in selecting the scoring criteria would have negative results on the scoring procedure. Teachers may find themselves unable to justify and even to control the scores.

In an essay-type test, there are many components that can be evaluated. The teacher should limit the number of components in order to control in a very precise way his scoring procedure. According to Madsen, one reason to evaluate a few factors at one time is that doing so helps us grade our papers more accurately and consistency. Another reason is to speed up our essay grading. A third reason for limiting the number of factors to be evaluate is to avoid unnecessary discouragement of our students. (Madsen .H.S. ; 1983:119)

The selection of the main factors that can be evaluated is an essential part in language testing. It may permit the examiner to conduct his scoring procedure in a more objective way. Such a decision may have three positive points. First, it helps the teacher to synchronise his judgements to some specific components. Second, the specification of the criteria of evaluation can speed up the scoring procedure, and makes it easy and practical in a mathematically precise way. Third, such a precision in limiting the number of components may have a positive effect on the students' attitudes towards the matter of evaluation. It may help them to pay more attention to what they are to be tested on. The criteria are usually of two kinds. The first includes all those criteria that focus on the accuracy of language use, such as grammar, mechanics, and diction. The second has been underscored by communicative approaches to language teaching, and includes criteria such as the relevance of content,

organisation of ideas, paragraphing, and style. These two types of criteria constitute the basis of any scoring procedure.

4. The Specification of the Criteria of Evaluation

There are seven criteria that should be taken as the main basis of evaluation. Each of them is divided into categories.

1. Relevance	
Categories	**Attributes**
A.The essay demonstrates superiority. It leads no doubt in the reader's mind that the student possesses a superior understanding of the question.	➢ Completely and directly answers the question. ➢ Supports the answer with detailed evidence that is correct. ➢ Provides insights not readily obtained from class courses. ➢ All information are correct.
B. The essay minimally answers the question.	➢ Demonstrates evidence of serious consideration of the topic. ➢ Provides some supporting data from class courses. ➢ Some information are correct.
C. The essay almost answers all or most part of the question.	➢ Supporting data is lacking, deficient, or incorrect. ➢ Part of the answer is wrong.
D. The essay answers less than half of the question.	➢ Provides no supporting evidence. ➢ Most of the answer is incorrect.
E- The essay is considered as a failing work	➢ There is a misinterpretation of the question. ➢ The response is completely off-topic whatever its writing quality. ➢ The reader is left with certainty that the student has not understood the question.

2. Paragraphing	
Categories	**Attributes**
A.The paper contains a clear, succinct and direct response to the question asked and develops that response through a sequence of reasonably ordered paragraphs. The essay has a discernible beginning, middle, and end.	➢ **The introduction** focuses the reader's attention on the subject of the essay in a thorough Paragraph of thought – evoking sentences leading effectively into the Thesis statement. ➢ **The thesis** clearly, specifically, and interestingly states or implies the main idea or ideas which the essay will explain or support. ➢ **The development** thoroughly supports and explains the thesis, or builds to a logical conclusion in a series of vivid interesting paragraphs. ➢ **The conclusion** logically completes the development of the thesis or build up to the main points of the essay.
B.The paper has minor weaknesses in paragraphing, but it contains evidence of the writer's ability to organise information into fluent and unified paragraphs.	➢ The paragraphing is limited in logical development of ideas into paragraphs. Both introduction and conclusion are lacking.
C.Serious difficulty in managing the tasks of the assignments. The essay lacks an overall plan with a beginning, middle and end.	➢ The paragraphs are somewhat disorganised. The key idea in paragraphs lacks development or illustration. They provide little evidence of the student's ability to develop an organised response

3. Expression	
Categories	**Attributes**
A. The essay demonstrates mastery of the elements of effective writing.	➢ The essay demonstrates an effective range of sentence clarity, variety and word-choice.
B. The essay reveals an acceptable mastery of the elements of effective writing.	➢ The essay displays sentence clarity, variety and generally appropriate choice. ➢ Some inappropriate words, which do not affect the quality of the essay. ➢ The meaning is comprehensible.
C.The essay demonstrates virtually no mastery of the elements of an effective writing.	➢ The essay demonstrates almost no sentence variety. ➢ The meaning of the sentence is ambiguous due to the consistently inappropriate or non-idiomatic word -choice. ➢ The writer's control of language may be imprecise, awkward, or clumsy.

4. Organisation & 5. Coherence	
Categories	**Attributes**
A . Ideas cogently developed.	➢ Ideas are logically organised and connected with clear transitions. ➢ The writer demonstrates clear understanding of the writing task. ➢ The writer has a good control of language to convey ideas with reasonable quality.
B. Ideas are somehow arranged.	➢ Ideas are adequately developed and organised, but they are not connected with transitions.
C. The essay is chaotic.	➢ Ideas are confusing and mishandled. There are no connections between ideas.

6. Grammar & 7. Mechanics	
Categories	**Attributes**
A.The essay demonstrates consistent mastery of language and mechanics.	➢ It is nearly free from errors of grammar and mechanics, and there is evidence of superior control of language.
B.The essay demonstrates competence .	➢ It is free from serious errors and is generally well-written and characterised by clarity.
C.The essay contains some errors.	➢ The essay contains some errors in grammar and mechanics, but not to continually distract the reader from the context.
D.The essay reveals lack of competence.	➢ The essay has serious and frequent problems in the use of language and sentence structure. ➢ It contains numerous errors in grammar, usage and mechanics that interfere with meaning. ➢ Such an essay suggests that the student is unable to deal competently with the question.
E. The essay demonstrates total lack of competence.	➢ It has severe and persistent errors in language and sentence structure. It contains a pervasive pattern of errors in grammar usage or mechanics that may interfere with readability giving the impression of distinctly inferior writing.

5. The Analytic Assessing Scheme

The third step that the teacher can do is to set up an analytic scheme based on the selected criteria of assessment. He may decide in advance on the precise basis for scoring. The marking scheme, according to Harrison (1990), should be carefully studied before any

attempt to test students' abilities. Harrison. A. A (1990: 111) sees that the marking scheme should be thought out at an early stage in the development of the test, since it is in principle a forecast of what the students will produce and so affects what is to be included in the assessment.

Our judgements of students' responses, whether they are formal or informal right or wrong, are mainly based on personal opinions. It is quite possible to make such judgements sound objective. The possibility to reach such an aim is to specify the criteria of evaluation in a scoring framework, which should be tried out on a trial version so that any necessary judgements can be made. This assessing system should be checked at least by another teacher as recommended by Basanta B. P. C (1995: 57), the marking criteria should be set before hand and candidates must be informed as how they will be scored.

Students are assumed to know how they will be evaluated. Such an intention may permit them to be more aware of the type of test they are supposed to respond to. Each student, according to Heaton J.B (1975:137), is able to see how his particular grade has been obtained. He is in a position to judge the quality of his work by himself. Moreover, each student will be encouraged to adapt his writing product to the task of the assignment without fear of being penalised for something he does not know. The specification of an analytic assessing scheme can be also maintained by a clear description of its elements. Each element must be studied and controlled by the specification of its attributes that it is based on.

6. The Selection of the Test Item

In any consideration of educational testing, a distinction must be drawn between the rather classroom formal tests and those informal ones. The selection of the essay test question is as important as the assessment procedure itself. It is not possible to look for objectivity in a test which can automatically increase subjectivity. An essay-type test is subjective in the sense that the question itself can have many interpretations. Such interpretations differ from

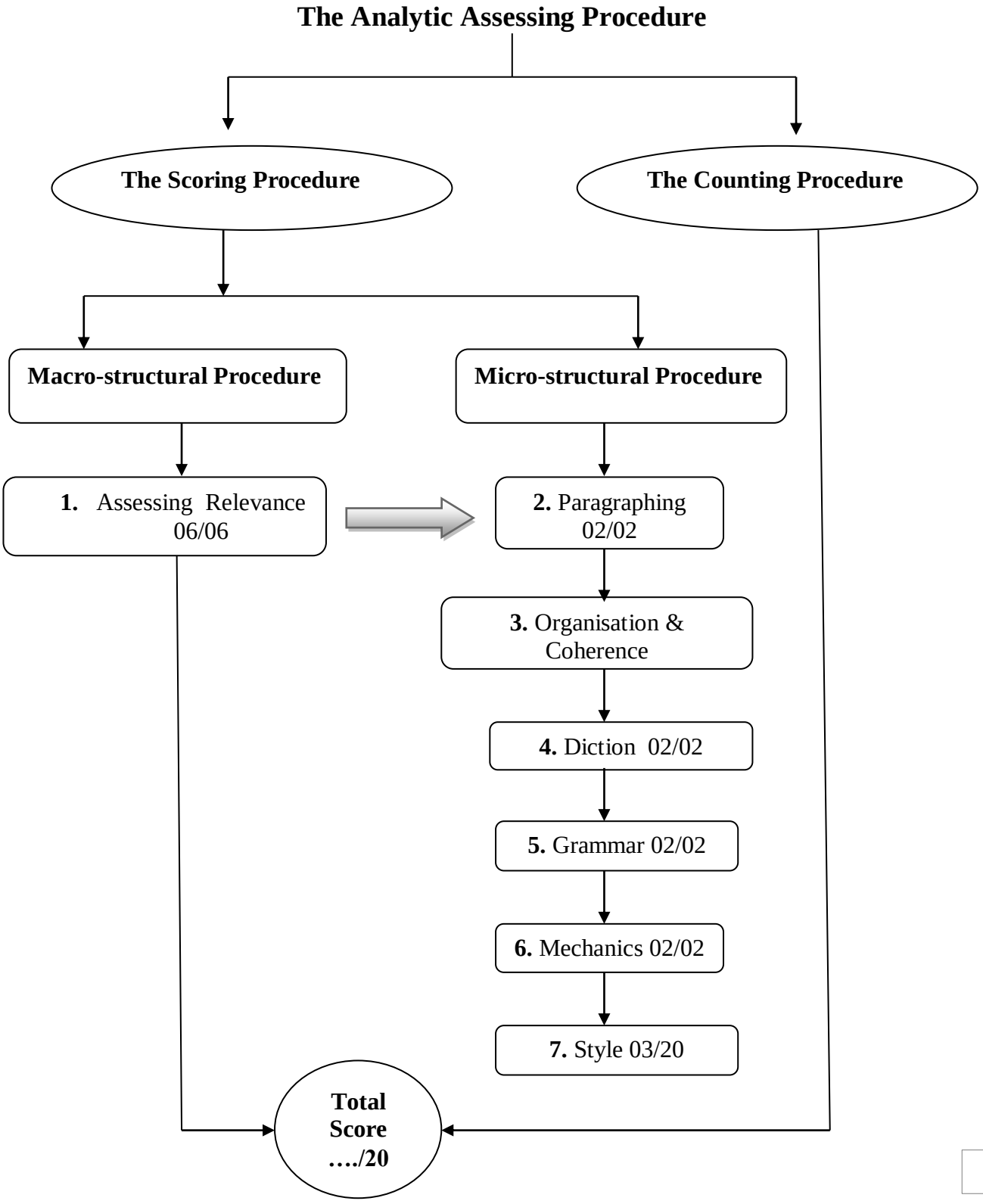
student to another and even from occasion to another. There is unreliability in the interpretation of the test question before it can result in the scoring procedure. The main difference between objective and subjective tests lies in the administration of the test items.

Objective tests are supposed to be more scientific than subjective tests because their results are always reliable. Such reliability is the result of the quality of the test which is easily understood and easily scored. Students can estimate in advance which grade they may obtain. They may feel secure not because the test can be easy, but because they can know about what they might write and what they may be asked to do. Such a feeling may permit the classroom teacher to be more dependent upon a consistent assessing scheme which can control his scoring procedure.

Before any attempt to write the test question, the classroom teacher should be aware of each word he is going to set forth. The analytic scoring scheme can be used to control the scoring system in a way that makes scores more reliable. However, in order to achieve such probability, the test question should be controlled and studied word by word. It must be adapted to the assessing scheme that should be given to the students before administering the test question. Furthermore, the test content should be introduced not to challenge the students but to give them meaningful feedback of various aspects of their performance. It should be prepared to the whole population not just to some of them. Students cannot know in advance which material their group will be tested on. They may feel insecure when they find themselves tested on something they do not know and with a type of material which is unfamiliar to them. They should be informed of how to respond to the same question that they are supposed to respond to in the test session. The analytic scoring procedure is a test in itself. It helps the teacher to select the test items that can accompany it in order to achieve formal and conclusive results.

7. The Analytic Scoring Procedure

The analytic scoring procedure consists of scoring in categories such as the relevance of content, paragraphing, organisation of ideas, diction, grammar, mechanics and style allotting each a mark and awarding total score out of twenty. Such a procedure is divided into two interrelated procedures: the scoring procedure and the counting procedure.



It is possible to use the "*Process Approach"* (Mahili, 1994) which involves multiple-drafts focussing on both content and language at separate stages. The starting point is by the scoring procedure, which is, in turn, divided into two procedures: Macro-structural and Micro-structural procedures.

8. An Analytic Assessment of an Essay Example

In order to make things less abstract, the Analytic Assessing Procedure is used with the following example of a Linguistics essay test.

The essay question:

'As opposed to Microlinguistics, Sociolinguistics has been established as a multidisciplinary field of investigation with a distinct approach. Analyse.

The essay answer

The answer is based on two main ideas:

- Microlinguistics is an independent study whereas Sociolinguistics is a multidisciplinary field of investigation
- Sociolinguistics adopts a distinct approach which is quite different from Microlinguistics approaches.

The answer must contain the following ideas:

1. **Microlinguistics** is seen as an independent discipline. It studied language without any reference to other approaches. It set up its own methods which are not shared with the other disciplines. (**01pt)**

2. **Sociolinguistics** as a new field of investigation is mainly dependent upon four disciplines such as Generative Grammar, Dialectology, Anthropology, and Sociology. **(01pt)**

3. Microlinguistics Approaches:

A-Structuralism: Language is a static phenomenon. It does not change. **(01pt)**

B-Generativism: Language is a dynamic phenomenon. It is generated by finite abstract rules (linguistic competence). **(01pt)**

4. Sociolinguistic Approach:

A-Language is a flexible phenomenon. It is subject to variation either cultural or social.**(01pt)**

B-Language is affected by the environment which imposes the rules of speaking. **(01pt)**

I- The Scoring Procedure

A. Macro-structural Scoring Procedure

- **Scoring the Relevance of Content**

The Macro-structural procedure is the first step that the classroom teacher starts with. It is a general evaluation of the relevance of content. It deals with the gist of the answer without any reference to the other criteria of evaluation such as paragraphing, organisation, diction, grammar, mechanics and style which are the main concern of the Micro-structural procedure. The classroom teacher makes use of his own answer to the subject-matter in order to compare it with those of the students. The answer is a set of ideas, and each idea is allotted a point. The amount of points constitutes the general mark allotted to the relevance of content.

As far as the relevance of the content is concerned, the possible way to control the mark given to the content is (1) to divide the content into main ideas and (2) to give each idea a specific point. The teacher makes a general reading of the whole answer. The twofold objective of such a whole reading are (1) to see whether or not the student writes on the subject, and (2) to search for the points of discussion that he is asked to deal with. The student's answer can be interpreted as either:

- knowledgeable,
- minimally answer the question ,
- some knowledge of the subject ,
- limited knowledge of the subject
- or misdirected (off-topic).

If the answer is acceptable the teacher starts by giving each idea a specific mark according to the essay answer outline he sets forth before administering the essay test .Such an answer is used as a model. For instance, he may allot 06/06 for an answer which satisfies the four attributes.

Such an answer is recognised as knowledgeable because the student possesses a superior understanding of the question. All the ideas, as a matter of fact, have been introduced in a very precise way. This leads no doubt in the reader's mind that the student has fully understood the question matter. In other cases, the student may not respond to the question with the whole ideas. He may minimally answer the question or the answer may contain just some or limited knowledge of the subject. In accordance with the essay answer the teacher may give for instance 04/06 to an answer which lacks two ideas if the exact answer should be based on six ideas. He may give 03/06 to an answer which lacks three ideas and 02/06 to an answer which lacks four ideas.

However, if the student's answer is formally wrong, the student will be given a mark agreed on in advance and set forth as misinterpretation of the subject. Such a mark is, according to the scoring scheme 01/20. The student's essay test is automatically dismissed. The response is completely off-topic whatever its writing quality is. The reader is left with certainty that the student has not understood the question. In such a situation, the micro-strutural procedure is not to be applied. The analytic assessing procedure starts with the content and ends with it whenever the message is deemed unacceptable.

B- Micro-structural Assessing Procedure

The second step in assessing the essay answer deals with the six other criteria which are chronologically organised according to the assessing procedure. The Macro-structural assessing procedure paves the way to the second step of assessment. The student's answer seems acceptable according to the teacher's knowledge of the subject. The Micro-structural assessing procedure proceeds by dividing the essay answer into six scoring criteria. Each of them is to be evaluated separately. Consequently, six procedures may take place as follow:

- **Assessing the Paragraphing of Ideas**

According to the teacher's knowledge of the subject, the answer should have a clear and

suitable organisation of ideas into appropriate paragraphs. The teacher forecasts in advance all the possibilities that may occur in organising the ideas of the essay answer. He tries to concentrate only on the organisation of ideas of the relevance of content into paragraphs without any reference to the other criteria such as, coherence, diction, grammar, mechanics and style. The main objective in doing so is (1) to ensure all the possibilities in organising the ideas that the students are likely to adopt , and (2) to limit the assessing procedure to just one aspect. The teacher knows in advance all the possibilities that can be used. For instance, the essay answer, provided above, consists of six ideas. These ideas can be organised into different paragraphs. There are, however, five possible ways to organise those paragraphs. The content of the essay can be organised in two, three, four, five or even six paragraphs as it is shown in the following tables:

1. Two paragraphs	**First Paragraph /** The first idea (1) with the second (2) **Second Paragraph /** The third idea (3 / A+B) with the fourth (4 /A+B)
	First Paragraph / The first idea (1) with the third (3 /A+B) **Second Paragraph /** The second idea (2) with the fourth (4 /A+B)

2. Three paragraphs	**First Paragraph /** The first idea (1) with the second (2) **Second Paragraph /** The third idea (3 / A+B) **Third Paragraph /** The fourth idea (4 / A+B)
	First Paragraph / The fourth idea (4 /A+B) **Second Paragraph /** The third idea (3 / A+B) **Third Paragraph /** The first idea (1) with the second (2)

3. Four paragraphs	**First Paragraph /** The first idea (1) **Second Paragraph /** The third idea (3 /+B) **Third Paragraph /** The second idea (2) **Fourth Paragraph /** The fourth idea (4 /A+B)
	First Paragraph / The third idea (3 /A+B) **Second Paragraph /** The first (1) idea **Third Paragraph /** The fourth idea (4 /A+B) **Fourth Paragraph /** The second idea (2)
	First Paragraph / The second idea (2) **Second Paragraph /** The fourth idea (4 /A+B) **Third Paragraph /** The first idea (1) **Fourth Paragraph /** The third idea (3 /A+B)
	First Paragraph / The fourth idea (4 /A+B) **Second Paragraph /** The second idea (2) **Third Paragraph /** The third idea (3 /A+B) **Fourth Paragraph /** The first idea (1)

4. Five Paragraphs	**First Paragraph /** The first idea (1) with the second (2) **Second Paragraph /** The third idea (3 /A) **Third Paragraph /** The third idea (3 /B) **Fourth Paragraph /** The fourth idea (4 /A) **Fifteh Paragraph /** The fourth idea (4 /B)

5. Six paragraphs.	**First Paragraph /** The first idea (1) **Second Paragraph /** The second idea (2) **Third Paragraph /** The third idea (3 /A) **Fourth Paragraph /** The third idea (3 /B) **Fifth Paragraph /** The fourth idea (4 /A) **Sixth Paragraph /** The fourth idea (4 /B)

N.b. It is sometimes possible to reverse the order of paragraphs.

The organisation of ideas is an important factor in responding to an essay question. It denotes the student's ability to put facts into a sequence of reasonably ordered paragraphs. The procedure to allot a mark to this category consists of the following steps: First, the teacher tries to count the number of paragraphs written by the students. As it is demonstrated above the number of paragraphs should signal the number of ideas. There are six ideas which can be built up into two, three, four, five or six paragraphs, according to the matter of discussion. Second, the teacher compares the ideas expressed by the students with the number of paragraphs.

- If the number of paragraphs signals the ideas of discussion, the student will be given a full mark (02/02). The essay should have a discernible beginning, middle and end.
- The student will be given 01/02 if the essay is somewhat loosely paragraphed. The written production has minor weaknesses in paragraphing, but it contains evidence of the writer's ability to organise information into fluent and unified paragraphs. The paragraphing is limited in logical development of ideas into paragraphs. There is, however, a serious trouble in introduction and conclusion. The student does not know how to introduce and end the ideas that logically focus the reader's attention into the thesis or interestingly complete the development of the thesis and build up to the main points of the essay.

- The student will be given 0, 5/02 if there is serious difficulty in managing the tasks of the assignments. The essay seems also to lack an overall plan with a beginning, middle and end. The paragraphs are somewhat disorganised, and difficult to decipher.

➢ **Assessing Organisation and Coherence**

Following the ideas of discussion and their organisation into unified paragraphs, the students are also judged for their ways to come to communicate such ideas in a more comprehensible and accurate writing framework. The scoring procedure deals with the three following attributes:

- The student will be given 02/02 if the ideas he introduces are cogently developed. They seem logically organised and connected with clear transition. There is also a good control of language to convey ideas with reasonable quality.
- The student will be given 01/02 if the ideas that he develops into paragraphs are well organised, but they are not connected with transitions.
- The student will be automatically penalised with 0, 5/02 if the essay seems chaotic confused and mishandled. There are also no connections between ideas.

➢ **Assessing Language Form (Diction, Grammar and Mechanics)**

Language form consists of the three interrelated components: vocabulary, grammar and mechanics. The assessing procedure is done at the same time. The three components are judged all together. The classroom teacher starts first by indicating the wrong words by symbols. Each symbol specifies the type of error as mentioned in the following table:

Type of Mistake	**Symbol**
The errors made in vocabulary or diction	V
The mechanical errors: • Punctuation • Capitalization • Spelling	 P C S

The grammatical errors:	
• Subject and Verb Agreement	AGR
• Verb Tense	VT
• Unwanted Word	X
• Word Missing	MW
• Word Order	W.O
• Word Form (noun-verb- adverb –adjective)	W.F
• Verb Form (gerund-participle)	V.F
• Article Misused	ART
• Preposition Misused	PREP
• Reference (unclear relationship between a pronoun and its antecedent)	REF

In the second step the teacher starts by counting errors. Marks are reduced or deducted from the general marks allotted to diction, grammar and mechanics. For instance, in grammar the student will be given a full mark (02/02) if the essay demonstrates consistent mastery of language. It is virtually free from errors of grammar, and there is evidence of superior control of language. The student is to be penalised for each error he will make. The substraction is to be applied whenever the student makes an error. For each error 0,5 is deducted from the general mark allotted to grammar. The student will be given 00/02 if he makes four errors. The same procedure is used with diction and mechanics.

- **Assessing the Style**

Assessing the style is to be left at the end of the scoring procedure. The style is a kind of conclusion the teacher draw after making use of the learners' results in language form in order to judge the quality of the answer. The assessing the style can be done as follow:

- The student will be given a full mark (03/03) if the essay is clear and free of errors. In other words, the essay displays mastery of elements of effective writing, and it exhibits an effective range of sentence clarity, variety of word-choice.
- The student will be given 02/03 if the essay is comprehensible, but some occasional errors in grammar, mechanics and diction affect the meaning of the sentence. For example the student will be given such a mark in style if he makes less than four errors in grammar, mechanics and vocabulary.
- The student will be given 01/03 if the essay has the same attributes as in model (2) in addition to some ambiguous words and sentences. The essay reveals virtually no sentence variety. The meaning of the sentence is ambiguous due to the inappropriate or non-idiomatic word-choice.
- The student will be given 0,5/03 if the essay is chaotic due to some frequent errors and ambiguity of some words and sentences. The writer's control of language may be imprecise, awkward, or clumsy. For instance, such a mark will be given to a style which contains more than four errors in grammar, mechanics and diction, in addition to the existence of some words and sentences which are incomprehensible and even unreadable.

II. The Counting Procedure

The scoring procedure permits the teacher to judge the quality of each scoring component by a mark. He counts all the obtained marks and mentions them in a scoring grid. For instance, a student may be given 07/20. The scoring grid may be used to permit the student to know exactly why he has been given such a mark as the example below:

Criteria	Scores
1. Relevance	03/06
2. Paragraphing	01/03
3. Organisation	01/02
4. Diction	0,5/02
5. Grammar	0,5/02
6. Mechanics	00/02
7. Style	01/02
Total Score	07/20

Such a mark (07/20) may reveal that the essay is weak. It has the following characteristics:

- The essay demonstrates some knowledge of the subject. It almost answers most part of the question. If the question needs six ideas of discussion, the student has demonstrated only three of them.
- The essay lacks an overall plan with the beginning, middle and end. There is serious difficulty in managing the tasks of the assignments. The paragraphs are somewhat disorganised. The key ideas in paragraphs lack development or illustration. They provide little evidence of the student's ability to develop an organise response.
- The essay is chaotic, confused and mishandled. There are no connections between ideas, which to some difficulties to understand the answer.
- The essay contains some errors in diction. There is an inappropriate or non-idiomatic word-choice.
- The essay suggests lack of competence. It has serious and frequent problems in the use of language and sentence structure. It contains numerous errors in grammar, usage and mechanics that interfere with meaning. Such an essay suggests that the student is unable to deal competently with the question.

- The meaning is comprehensible, but there are some sentences which are ambiguous and contain some occasional errors in diction, grammar and mechanics. The lack of language competence does not affect the essay but it downgrades its quality.

9. The Analytic Assessing Limitations

The analytic assessing procedure could be very effective in the sense that it assists the teachers identify in a high accurate manner their students' deficits in many writing areas, as it facilitates things to design the appropriate and effective feedback to set up and boost the learning of writing.

Yet, the analytic assessing process is not without limitations. We may point out to three major drawbacks. First, it is blamed to be time and effort consuming. An analytic approach to assess writing would certainly require the teacher to read and evaluate each single written paper at least seven times. He would read the paper to check up seven different components i.e. he is to read a first time to assess the relevance of the content, a second time to review the paragraphing, a third time to check the organisation, a fourth time to gauge diction, a fifth time to evaluate the grammar, a sixth time to weigh up the mechanics and a seventh time to estimate the style.

Second, and because of the high number of the EFL students in our universities, such a procedure would never attain high rates of effectiveness, objectivity, consistency and most importantly reliability. The assessment of the very first written papers would achieve a high degree of reliability however it would never be the case when the number of papers exceeds one hundred and more. The last disadvantage to consider is that such process would not be practical. The huge number of written essays to assess would never allow the teacher to determine accurately his learners' common deficits and hindrances. In other way, he would never have an obvious insight about the general lacunae of his learners, thus he would be unable to design a collective effective feedback as remedial work.

Consequently and in this respect, we would point out the possibility to overcome all those deficiencies mentioned above through transforming the assessing scheme suggested into a computer programme application. Such kind of assessment has existed since a long time ago. Moreover, a lot of the automatic assessing applications are available in the market of pedagogy.

10. Automatic Writing Assessment

Automatic writing assessment refers to the use of the computer technology to assess and evaluate the written production. It is computer software that is designed to aid the writing teachers in the process of assessing their students' essays. Such a technology is principally used to defeat time, reliability, and credibility matters in the assessment of the written language. A lot of surveys attempted to weigh up the exactness and consistency of the automatic essay assessment programme. The findings of numerous investigations pointed out a great conformity between the computer assessment and that of the human (Burstein 1999, Foltz 2003, Nichols 2004, and Page, 2003).

The use of the computer to assess the written language is not a recent issue. It is dated back to 1960s. However, it has not become a reality until the unprecedented innovation of the computer sciences in the 1980s. A group of researchers working for specific firms put forward computer programmes to assess the written language. They have asserted that their computer applications agree with human assessors to really great extents. Accordingly, computer assessing has been put into practice in a lot of important testing courses. In the United States of America, for example, the use of the computer was firstly introduced to assess the writing components of the Graduate Management Admissions Test (GMAT). Such a test, mainly used in the admission to a graduate management program, is now assessed by a teacher and by software rather than by two teachers and shows very elevated rate of reliability and consistency.

Automatic assessment to EFL learners' writings may be able to assess and determine in a mathematically precise way the students' lacunae and the extents to which they are improving their writing skills. Thus, things will be simplified as far as designing an effective feedback will be concerned. In other words, automatic assessment helps the teacher design an effective feedback for a great number of students. In this respect, Ruth Breese (2012) determines the aims of the feedback saying: "Although feedback is traditionally associated with lavish use of the red pen, it is important to remember that the main purpose is to provide a channel for teachers to communicate constructively with students and help them to develop as writers." Ruth Breese (2012:139)

It is obvious enough that the feedback is kind of facilitator or a vehicle through which the teacher may set up the whole process of teaching. Hattie and Timperley argue also that *an* effective feedback can have a major effect on both learning and achievement (2007:57). They postulated that the ideal feedback should include three responses: "Feed up, feedback, and feed forward" that is to say, feedback should provide the student with information concerning, the learning goals, student performance, and ways of improvement. Feedback is not only for learners to bridge the breach between the actual and the wanted skills. It is also vital for educators to assess, regulate, and augment their teaching exercise efficiency and success.

Finally, the present study highlights five automatic writing assessment applications namely the Project Essay Grader, the Intelligent Essay Assessor, E-rater, IntelliMetric, and Bayesian Essay Test Scoring System.

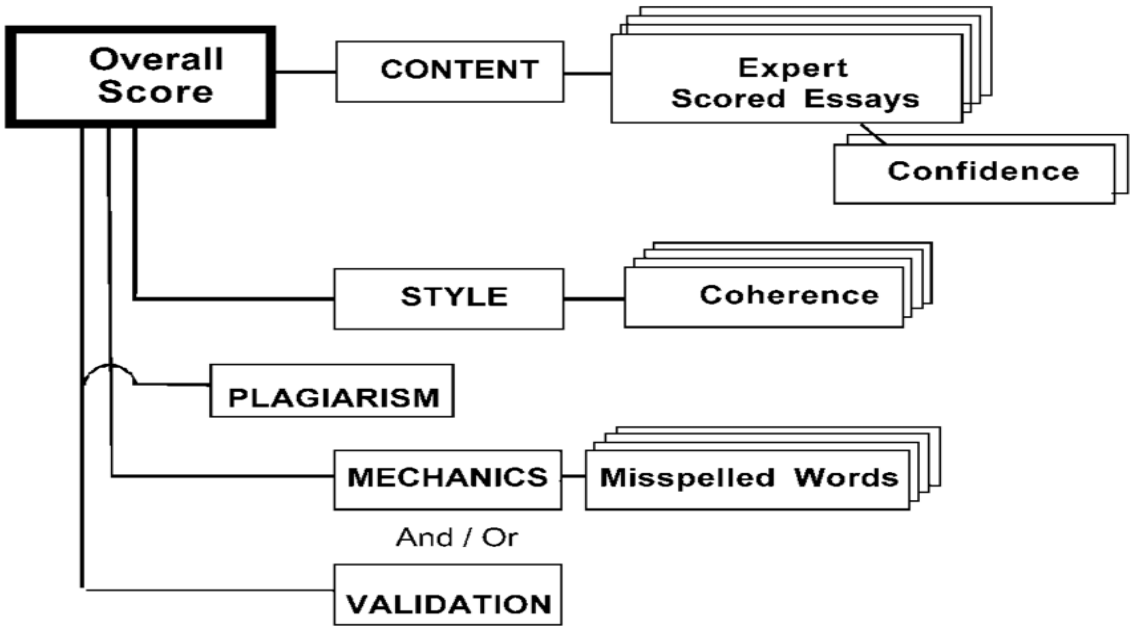
- **Project Essay Grader (PEG)**

Project Essay Grade was designed by Ellis Page in 1966 to achieve more practicality, reliability and effectiveness for the essay assessing process involving a large number of students. PEG makes use of association to foresee the intrinsic value of the essays. Page utilizes the item '*trins'* and '*proxes'* to illustrate the way PEG evaluating and assessing a written piece. 'Trins' refer to the basic components such as grammar, diction, fluency, and

punctuation. On the other hand, ‘proxes’ designates the connection of the intrinsic variables. Thus, ‘proxes’ refer the appropriateness of the use grammar rules, vocabulary, etc.

Project Essay Grader is very effective in allocating marks that are similar to those of human assessors. Besides, the application can computationally determine and identify the writing inaccuracies committed by the learners. Nevertheless, PEG was condemned for paying no attention to the semantic facet of the written language and spotlighting more the exterior or the surface structure (Chung & O’Neil, 1997). Simply put such an assessing machine fails to identify the content related aspects of a written piece or an essay i.e. its relevance, organization, style paragraphing and etc. The Project Essay Grader does not supply writing feedback to learners. It was criticized to be weak as far as rating exactness as concerned. It was possible to trick the programme through developing very long essays since it was based upon an indirect assessment of the writing skill.

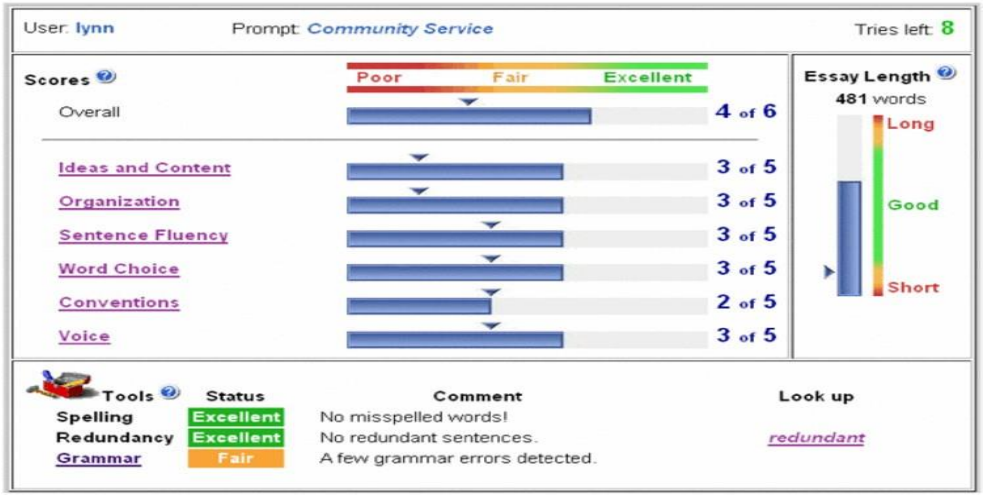
- **Intelligent Essay Assessor**



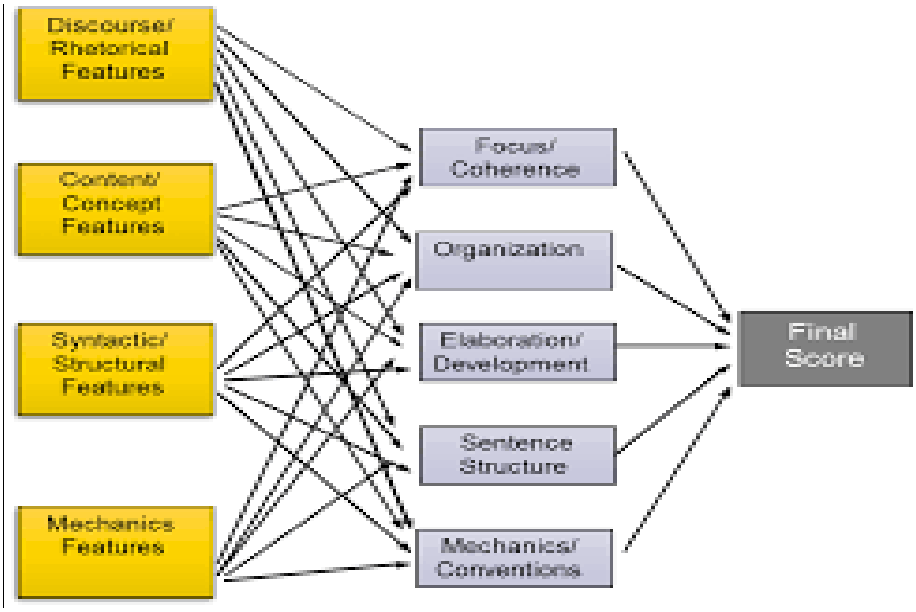
Intelligent Essay Assessor Architecture

Intelligent Essay Assessor, IEA examines and assesses writing by means of a semantic essay-analysis method. Such a method originated the 'Latent Semantic Analysis' approach put forward by the following psychologists Thomas Landauer, Peter Foltz and Darrell Laham.

Foltz sees 'Latent Semantic Analysis', LSA as "*a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information*" (Foltz, 1996:2). The function of Intelligent Essay Assessor involves is very simple. In order to weigh up the general quality of an essay, it necessitates to be trained on well written texts. Then, the written essay needs to be typified through vectors as mathematical representation of the essay. Finally, the theoretical relevance of the content of the targeted essay is compared to other texts (Landauer et al., 1998). The Intelligent Essay Assessor considers much the content related characteristics rather than the form related ones. But it does not mean that the Intelligent Essay Assessor disregards formal aspects such as grammar and mechanics. In other words, even though the IEA applies an LSA approach to assess chiefly the quality of the content of an essay, it also provides feedback on grammar, style and mechanics as it is clearly noticeable in the following feedback figure generated by the so-called Intelligent Essay Assessor.



- **IntelliMetric**



IntelliMetric Architecture

IntelliMetric makes use of three main computer technologies: artificial intelligence (AI), natural language processing (NLP), and statistical technologies. IntelliMetri is conceived to adopt the human wisdom. In other words, it is developed in a way to comprehend natural language thus being able to assess any written language. Consequently, IntelliMetric deals with the essay according to the key features of standard written English. IntelliMetric is made able to store each rate related to specific characteristics in an essay answer. It is also claimed that the scoring system "learns" the characteristics that human raters likely to value. However, IntelliMetric needs to be trained with a set of pre-assessed written essays with marks allocated by writing teachers. These samples are used as a basis to deduce the assessing scale and the human assessing attitudes as well.

IntelliMetric is made able to assess about 300 semantic, syntactic, and discourse components in a written piece through the use of Artificial Intelligence and Natural Language Processing technologies. To put it simple, IntelliMetric is conceived to simulate the human brain. Thus it may imitate the human method in assessing the essay. It is based upon a difficult scheme of information processing. IntelliMetric works in many dimensions. It is recursive. It uses different assessments based on binary system.

Moreover, such an assessing tool may assess essays written in many languages such Dutch, French, German, Italian, Arabic, Japanese, Spanish and English.

- **Bayesian Essay Test Scoring System**

The Bayesian Essay Test Scoring System is suggested by Lawrence M. Rudner. BETSY is somehow different from the previous tools since it may be recognized as both an assessing and a research tool together.

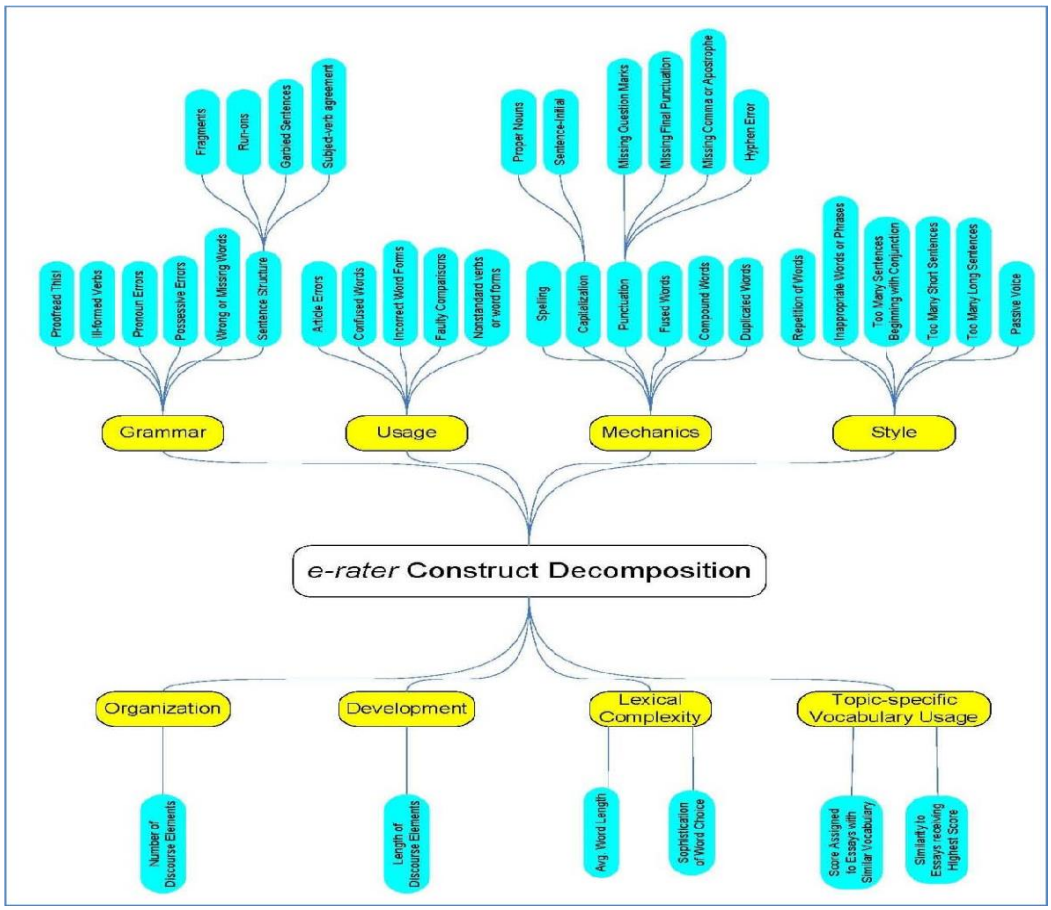
Such a system is based on the Bayesian theorem approach. Bayesian approach have manufactured many programme such as recognizing spam and other undesired e-mails based on their similarity with previously classified e-mail.Bayesian is available under two samples used in assessing texts: the Multivariate Bernoulli sample and the Multinomial sample. The former assesses the existence of specific components in a written paper. The latter verifies the various use of a definite component in an essay. The Multinomial form works out quite quickly compared to the Bernoulli form.

The Bayesian approach uses key concept such as stop words, feature selection, and stemming. Stop words points out prepositions, pronouns, adjectives, and various articles. Stemming represents the procedure of removing suffixes to find stems. For example, obtaining “form” as a stem for formation, transformation, transformational, formative and formed. Feature selection, however, indicates reducing of randomness and disorder. BETSY requires a training of 1000 written samples to grab how to assess new papers.

During the past four decades, numerous investigations took place wishing to highlight methods to combine and apply the computer technology to the writing assessment. More recently, progressively more innovative micro processor technology has allowed writing skills to be assessed automatically. In this respect, we can mention the highly developed researches of Rudner & Gagne, 2001; Rudner & Liang, 2002; Hamp-Lyons, 2001; and Attali & Burstein, 2003. Such systems of assessments do not usually assess straightly the basic components of a written essay as teachers do, but they rather employ correlations of the basic components to

envisage the rate of an essay. The automatic assessing essay systems cited above employ a wide range of procedures to design an instant comment, feedback, and mark. Finally, each automatic assessing architecture request different amounts of written essays to train the system.

- **E-rater**



The E-rater or the electronic essay rater was meant to determine the linguistic components in a written text. It was established by the Educational Testing Service (Educational Testing Service is an educational testing and assessment organization in USA. It has put forward many famous tests such as TOFL). E-rater exploits the natural-language processing method. The E-rater was conceived to spot definite lexical and syntactical hints and prompts in a written paper (Burstein, 2003; Kukich, 2000).

E-rater employs a corpus-based approach. E-rater is trained on a corpus of written works assessed previously by at least two teachers on a 6-mark scale to generate a sample, in which actual essay data are used to examine sample essays. The architecture of the E-rater encompasses three elements mainly the syntactic, the discourse and the topic-analysis. Those three elements generate the final assessment of the essay. Each component considers a specific area. Consequently, the syntactic element considers the use of the syntactic structures such the use of the clauses. As for the discourse component deals with linking words and conjunctions to identify how well the essay is organised. Finally, the topical analysis element spots lexis usage and topical content.

11. Conclusion

The essay-type test, as it is always referred to, is subjective in the sense that it requires more personal judgements rather than objective interpretations. Such subjectivity is still a problematic which is the result of individual standards of the grader and unreliability of scoring procedures. However, in order to put an end to such a problematic some guidelines have been suggested as a tentative solution to the formal scoring of an essay test question.

The assumption that scores unreliability is caused by faulty scoring procedure has lead the classroom teacher to be more aware about the extraneous factors which are closely related to the measurement setting and to variations in personal characteristics rather than to the outcome being measured. To the extent that the scoring procedure is free of such inappropriate influences, the results (and decision based on the results) will be consistent and stable.

Accordingly, reliability can be increased if we standardize the assessing procedure. The proposed guidelines are set forth as testing measurements for practical purposes. They are not supposed to have attained objectivity in its exact sense or even a substitute for it such as probability .We constantly fall into the same error of supporting that an absolute solution to

the scoring of assays is attainable and we look for revelation. Every technique in testing written responses remains tentative forever, but we shall not cease from exploration. We must admit that the end of our exploring will be to arrive where we started and know the place for the first time.

References

Books

Anatasi, A. (1982). *Psychological Testing*. (Fifth edition).London: Collier Macmillan

Bachman, L.F (1990). *Fundamental Consideration in Language Testing.* Cambridge: CUP

Bachman, L.F, and Palmer, A.S. (1996). Language Testing In Practice: Designing and developing useful language test. Oxford: OUP.

Brown, D.H. (1987).*Principles of language Learning and Teaching.* Second Edition. Prentice Hall Regents, Englewood Cliffs

Brumfit, C.J &Johnson (1991). *The Communicative Approach to Language Teaching.* Oxford: OUP

Carter, V.Good, (1973*) Dictionary of Education*. New York: MC Graw-Hill Inc

Charles, A.D. (1997).*Holistic Scoring Methods.* University of Texas.

Corder, S. Pit. (1985). *Introducing Applied Linguistics*. New York: Penguin Books Ltd

Cronbach, L. J. (1984).*Essentials Of Psychological Testing.* Fourth Edition. New York: Harper Andrew.

Dickens, P.R. and Germaine, k. (1983). *Evaluation.* Oxford: OUP

Doff, A. (1995). *A Training Course for Teachers.* Cambridge: CUP

Gronlund, N.E. (1985).*Measurement and Evaluation in Teaching*. Fifth Edition. New York: Macmillan

Hamp-Lyons, L and Heasley, (1987).*A Course in Writing English For advanced and Professional Purposes.* Cambridge: CUP

Harris, D.P. (1969).*Testing English as a Second language*. New York: McGraw-Hill Ltd

Harris, J. (1993).*Introducing Writing.* London. Macmillan Ltd

Harrison, A.A. (1990).*A Language Testing Handbook .London*: Macmillan Ltd

Heaton, J.B. (1975).*Writing English Language Tests.* London: Longman group Ltd

Lado, R. (1962). *Language Testing*: The Construction and Use of Foreign Language Tests. New York: McGraw-Hill Company

Littlewood, W. (1981) *Communicative Language Teaching*, London: Cambridge: CUP

Madson, H.S. (1983). *Techniques in Testing.* Oxford: OUP

Nitko, A.J. (1983).*Educational Tests and Measurement*: An Introduction. New York: Harper Andrew

Nolasco, R and Arthur, L. (1983). *Large Classes*. London: Macmillan Ltd. Cambridge: CUP.

Nunan, D. (1989). *Designing tasks for communicative classroom*. Cambridge: CUP.

Raimes, A. (1998). *Teaching Writing.* New York: Oxford University Press.

Ruth, B. (2012). *Rethinking academic writing pedagogy for the European university.* Amsterdam: Editions Rodopi B.V.

Underhill, N. (1990).*Testing Spoken Language*: Handbook of Oral Testing Techniques. New York: Mc Crow-Hill

Ur, P. (1997).*A course in language teaching*: Cambridge teacher training and development. Cambridge: CUP.

Weigle, S.C (2002).*Assessing writing*: Cambridge /CUP.

Widdowson, H.C (1978). *Testing Language as communication*, London: Oxford University Press.

Wilkins, D.A (1976). *Notional Syllabus*, London: Oxford University Press.

Yorkey, R.C. (1982).*Study Skills for Students of English*. Second Edition. New York: McGraw-Hill.

Articles

Attali, Y. & Burstein, J. (2006). *Automated Essay Scoring with e-rater.* V.2. Journal of Technology, Learning, and Assessment (JTLA), 4(3).

Basanta, C.P. (1994). *Coming to Grips with Progress Testing*. ETF, 33, 3, PP. 55-58

Burstein, J. (2003). *The e-rater scoring engine: Automated Essay Scoring with natural language processing.* In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J. & Chodorow, M. (1999, June). *Automated Essay Scoring for non-native English speakers.* Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Charney, D. (1984). *The Validity of Using Holistic Scoring to Evaluate*. Writing. Research in the teaching of English, 18, 65-81.

Chung, K. W. K. & O'Neil, H. F. (1997, April). *Use of networked collaborative concept mapping to measure team processes and team outcomes.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Clark. John, L.D. (1983).*Language Testing: Past and Current Status-Direction for the Future.* The Modern Journal, 67:31-43.

Davies, A (1978).*Language Testing*. In Language Teaching and Linguistics: Abstract, 11, 3-4.
Godshalk, F.L., Swinford.F. and Coeffman.W.E. (1996). *The Measurement of Writing Ability.* ETS Research Monograph 6.Princeton, N J: Educational Testing Service.

Hattie, J. Timperley, H. (2007). *The power of feedback*. The American Educational Research Association.

Ingram, E. (1970).*Attainment and Diagnostic Testing. In Davies (ed)* "Language Testing Symposium". Oxford: OUP.

Kenji, K and Kathleen, S. (1999).Testing Writing .Lancaster University Press

Mahilli. (1994). *Responding to Student Writing*. ETF, 32, PP. 24-27

Morrow, K. (1979).*Communicative Language Testing*: Revolution or Evolution? In Brumfit (ed.).1979, "The Communicative Approach to Language Teaching". Oxford: OUP

Nichols, P. D. (2004, April). *Evidence for the interpretation and use of scores from an Automated Essay Scorer.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.

Page, E. B. (2003). *Project Essay Grade: PEG.* In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pelliner, A.E.G. (1970).*Subjective and Objective Testing.* In Davies (ed.) "Language Testing Symposium":19-35

Pollit, A. (1990).*Giving Students a Suporting Chance: Assessment by Counting and by Judging.* In "Language Testing in the Nineties", (ed.) J.C. Alderson and B. North. Oxford: OUP.

Raatz, U. (1981).*Are Oral Tests Tests?* .In Klein-Braley and Stevenson, 1984:197-212.

Raimes, A. (1981).*Composition: Controlled by the teacher, free for students.* ETF, 19, 3, pp.2-7

Rudner, L. & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer. (*ERIC Digest number ED 458 290).

Wiseman, S. (1949).*The Marking of English Compositions in Grammar School Selection*. In British Journal of Educational Psychology.19, 200-209

Turkish Online Journal of Distance Education-TOJDE January 2006 ISSN 1302-6488
Volume: 7 Number: 1 Article: 5

Automated Essay Scoring

Semire DIKLI
Florida State University
Tallahassee, FL, USA

Abstract/

The impacts of computers on writing have been widely studied for three decades. Even basic computers functions, i.e. word processing, have been of great assistance to writers in modifying their essays. The research on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004). AES is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). Revision and feedback are essential aspects of the writing process. Students need to receive feedback in order to increase their writing quality. However, responding to student papers can be a burden for teachers. Particularly if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student essays might be quite time consuming. AES systems can be very useful because they can provide the student with a score as well as feedback within seconds (Page, 2003). Four types of AES systems, which are widely used by testing companies, universities, and public schools: Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater, and IntelliMetric. AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The accuracy and reliability of these systems have been proven to be high. The search for excellence in machine scoring of essays is continuing and numerous studies are

being conducted to improve the effectiveness of the AES systems.

Keywords: Assessment, Writing, Feedback Mechanism, Assistive Technologies

Introduction

The impacts of computers on writing have been widely studied for three decades. Even basic computers functions, i.e. word processing, have been of great assistance to writers in modifying their essays. The research on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004). AES is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). Revision and feedback are essential aspects of the writing process. Students need to receive feedback from the teacher in order to increase their writing quality. However, responding to student papers can be a burden for teachers. Particularly if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student essays might be quite time consuming. AES systems can be very useful because they can provide the student with a score as well as feedback within seconds. Also, the scores would be much more descriptive than the ratings provided by two human raters (Page, 2003).

Machine scoring technologies can also increase the practicality in administering large- scale assessments of writing ability (Bereiter, 2003). Employing human raters could be quite expensive in terms of time and resources. It is necessary to include more than one rater in large-scale writing assessments to reduce the bias the individual scorers might have. The training of multiple raters on a holistic scoring rubric is necessary but costly as well. In this case, it might be cost-effective to use an AES system (Bereiter, 2003; Chung & O'Neil, 1997; Page, 2003). Besides being a time-and money-saver, automated essay scoring systems

are claimed to provide variety in feedback, not only on grammatical issues, but also on discourse related issues (Shermis & Burstein, 2003, p. xiv). Myers (2003) claims that this reduces not only the teacher's paper load, but also the issues of concern (e.g., subjectivity) with teacher assessment. Similarly, Hamp-Lyons (2001) highlights the advantages of AES technology as follows, the ability to perform repeated functions without boredom and variation, adaptability (within preprogrammed pathways), flexibility (testing can be carried out at any time, for a range of purposes, and on any number of candidates), and the ability to make decisions without being judgmental (in the sense of being biased) or confrontational (p. 121).

Moreover, Page (2003) states that "the automated ratings would surpass the accuracy of the usual two judges. (Accuracy is defined as agreeing with the mean of judgments)" (p. 46). Finally, providing "a third voice" (p.15) about student writing, these types of programs can be effective tools in student-teacher conferences (Myers, 2003). A number of studies are conducted to prove the accuracy and reliability of the AES systems with respect to the writing assessment and the agreement rate between human raters and AES systems are found to be high (Attali, 2004; Burstein & Chodorow, 1999; Elliot, 2000a, 2000b, 2001c, 2002, 2003b, 2003c; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003, 2004).

Computerized scoring has many weaknesses as well. Hamp-Lyons (2001) stressed the lack of human interaction as well as the sense of the writer and/or rater as person. Similarly, Page (2003) stated that the computers could not assess an essay as human raters do because the computer would do "what it is programmed to do" and it wouldn't "appreciate" an essay (p. 51). Another

criticism is the construct objections. That is, the computer counts variables that might not be "truly" important in essay grading, i.e., focusing on formal aspects rather than organizational ones (Page, 2003; Chung & O'Neil, 1997).

Automated Essay Scoring Systems (AES)

Four types of AES systems are widely used by testing companies, universities, and public schools. The first one is Essay Grade (PEG), which is known as the first AES system built in AES history (Kukich, 2000; Rudner & Gagne, 2001; Page, 2003). The second one, Intelligent Essay Assessor (IEA), is developed by Landauer, Laham, and Foltz using Latent Semantic Analysis (LSA) features (http://lsa.colorado.edu/whatis.html). Another AES system, E-rater, has been used by the ETS (Educational Testing Service) to score essay portion of GMAT (Graduate Management Admissions Test). The final AES system is called IntelliMetric. It is developed by Vantage learning and used by the College Board for placement purposes (Myers, 2003).

Project Essay Grader (PEG)

Project Essay Grader (PEG) was developed by Ellis Page in 1966 upon the request of the College Board, which wanted to make the large-scale essay scoring process more practical and effective (Rudner & Gagne, 2001; Page, 2003). PEG uses proxy measures to predict the intrinsic quality of the essays. Proxies refer to the particular writing construct such as average word length, essay length, number of semicolons or commas, and so on (Kukich, 2000; Chung & O'Neil, 1997; Rudner & Gagne, 2001).

One of the strengths of PEG is that the predicted scores are comparable to those of human raters. Second, the system is computationally tractable. In

other words, it is able to track the writing errors made by the users. Next, its scoring methodology is straightforward. PEG contains a training stage and a scoring stage. The system is trained on a sample of essays in the former stage. In the latter stage, proxy variables are determined for each essay and these variables are entered into the prediction equation. Finally, a score is assigned by computing beta weights from the training stage (Chung & O'Neil, 1997). PEG has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures (Kukich, 2000; Chung & O'Neil, 1997). Failing to detect the content related features of an essay (organization, style etc.), the system does not provide instructional feedback to the students. Also, an early version of the system was found to be weak in terms of scoring accuracy. The main concern was the vulnerability of the system to cheating. Since PEG used indirect measures of writing skill, it was possible to trick the system, i.e., writing longer essays (Kukich, 2000). PEG was modified on several aspects in 1990s. It incorporated not only several parsers and various dictionaries, but also special collections and classification schemes (Page, 2003; Shermis & Barrera, 2002).

Intelligent Essay Assessor (IEA)

Another AES system, Intelligent Essay Assessor (IEA), analyzes and scores an essay using a semantic text analysis method called Latent Semantic Analysis (LSA) (Lemaire & Dessus, 2001). LSA approach was created by psychologist Thomas Landauer, a psychology professor at the University of Colorado at Boulder, with the assistance of Peter Foltz, a professor at the New Mexico State University and Darrell Laham, a PhD student at UC (Murray, 1998). IEA is produced by the Pearson Knowledge Analysis Technologies (PKT) (Psotka & Streeter, (n.d.); http://www.knowledge- technologies.com). A richer description of LSA and IEA is

provided below.

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is defined as "a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information" (Foltz, 1996, p. 2). LSA first processes a corpus of machine-readable language and then represents the words that are included in a sentence, (http://lsa.colorado.edu/whatis.html) paragraph, or essay. LSA measures of similarity are considered highly correlated with human meaning similarities among words and texts. Moreover, it successfully imitates human word selection and category judgments (Landauer, Laham, & Foltz, 2003). The underlying idea is that the meaning of a passage is very much dependent on its words and changing even only one word can result in meaning differences in the passage. On the other hand, two passages with different words might have a very similar meaning (Landauer, Laham, & Foltz, 2003). The underlying idea can be summarized as "meaning of word1+ meaning of word 2 + ……+meaning of word n = meaning of passage" (Landauer, Laham, & Foltz, 2003, p. 88). The educational applications of LSA include picking the most suitable text for students with different levels of background knowledge, automatic scoring of essay contents, and assisting students in summarizing texts successfully (http://lsa.colorado.edu/whatis.html).

In order to evaluate the overall quality of an essay, LSA needs to be trained on domain- representative texts (texts that best represent the writing prompt). The essay, then, needs to be characterized by LSA vectors (a mathematical representation of the essay). Finally, the conceptual relevance and the content of the essay are compared to other texts. When compared to content related factors

(e.g., argument, comprehensibility, style), mechanical and syntactic features are easier to separate from other factors. The reason is that content related factors are very much affected by the word choice. Previous research on automated essay scoring has concentrated on the analysis of style. Unlike other methods, the emphasis of LSA is on the conceptual content of an essay (http://lsa.colorado.edu/whatis.html; Foltz, Laham, & Landauer, 1999).

In the LSA based approach, the text is represented as a matrix. Each row in the matrix stands for a unique word, while each column stands for context. Each cell involves the frequency of the word. Then, each cell frequency is considered by a feature that denotes not only the importance of the word in that context but also the degree to which the word type carries information in the domain discourse (http://lsa.colorado.edu/ whatis.html). The semantics of a word is verified through all the contexts that the word occurs. The number of occurrences of each word in a text determines the semantic space. For example, 300 paragraphs and 2000 words provide a 300X 2000 matrix. Here, while each word is represented by a 300-dimentional vector, each paragraph is represented by a 2000-dimentional vector. By reducing these dimensions, LSA induces semantic similarities between words. This reduction is critical since it permits the representation of the word meanings through the context in which they occur. The number of dimensions is also crucial. That is, if the number is too small, much of the information will be lost. On the contrary, if the number is too big, limited dependencies will be drawn between vectors. According to this method, the semantic information is determined only through the co- occurrence of words in a large corpus of texts (Lemaire & Dessus, 2001).

Intelligent Essay Assessor (IEA)

It is claimed that unlike other AES systems, IEA's main focus is more on the content related features rather than the form related ones. However, this does not mean that IEA provides no feedback on formal aspects, i.e., grammar and punctuation, in an essay. In other words, even though the system is uses an LSA based approach to evaluate mainly the quality of the content of an essay, it includes scoring and feedback on grammar, style and mechanics as well (Landauer, Laham, & Foltz, 2000; Landauer, Laham, & Foltz, 2003; Streeter, Psotka, Laham, & MacCuish, 2004).It is also claimed that IEA can successfully analyze not only the content-based essays, but also the creative narratives. The system needs to be trained on a set of domain- representative texts in order to judge the overall quality of an essay. For example, a biology book can be used to evaluate a biology essay. IEA uses three methods to analyze an essay:

Y pre-scored essays of other students,

Y expert model essays and knowledge source materials,

Y internal comparison of an unscored set of essays" (Landauer et al., 2003, p. 90)

These methods allow IEA to compare the student essay with similar texts in terms of the content quality (Landauer et al., 2000; Landauer et al., 2003; Streeter et al, 2004). IEA, first, compares the content similarity between the student essay and other essays on the same topic that are scored by human raters and determines the closeness between them (Landauer et al., 2000; Rudner & Gagne, 2001; Streeter et al, 2004). It, then, predicts the overall score by adding "corpus-statistical writing-style" and mechanics (Hearst, 2000, p. 28). It also spots plagiarism and provides feedback (Landauer et al, 2000; Landauer et al., 2003). As part of the usual procedure of IEA, each essay is compared to every other in a set. The essays that

are extremely similar to each other are examined by LSA. Regardless of substitution of synonym, paraphrasing, or rearrangement of sentences, the two essays will be similar with LSA (Landauer et al., 2003). Detecting plagiarism is an essential feature since this type of academic dishonesty is quite hard to detect by human raters, particularly when grading large number of essays (Shermis, Raymat, & Barrera, 2003).

Landauer, Laham, and Foltz (2000) point out the basic technical difference between IEA and other AES systems as follows:

Other systems work primarily by finding essay features they can count and correlate with ratings human graders assigned. They determine a formula for choosing and combining the variables that produces the best results on the training data.

They then apply this formula to every to-be-scored essay. What principally distinguishes IEA is its LSA-based direct use of evaluations by human experts of essays that are very similar in semantic content. This method, called vicarious human scoring, lets the implicit criteria for each individual essay differ (p.28). The producers of IEA, Pearson Knowledge Technologies (PKT), report that they benefited from the system greatly since it needs smaller numbers of pre-scored essays to train. Unlike other AES systems, which require 300-500 training essays per prompt, IEA only requires around 100 pre-scored essays (http://www.knowledge-technologies.com; Landauer et al., 2003).

Another reason is that IEA does not require a representative sample of all scores in the rubric, either. They claim that the system is so intelligent that it can determine the scale of the essay. For example, the system is able to predict what an essay with 6 point looks like in a 6 point holistic scale without seeing large numbers of essays with 6 point (http://www.knowledge-technologies.com).

Finally, the developers of IEA claim that the system does not evaluate the creativity and reflective thinking. It, however, assesses "expository essays on factual topics", i.e., description of a psychological theory, the function of the heart (Murray, 1998). IEA's future plans include moving from global assessment features such as flow and coherence to more specific ones such as the voice and the audience (Landauer et al., 2003).

E-Rater and Criterion

The Electronic Essay Rater (E-rater) was developed by the Educational Testing Service (ETS) to evaluate the quality of an essay by identifying linguistic features in the text (Burstein & Marcu, 2000; Burstein, 2003). E-rater uses natural language processing (NLP) techniques, which identify specific lexical and syntactic cues in a text, to analyze essays (Kukich, 2000; Burstein, 2003). A detailed description of natural language processing and information regarding the structure and functions of e-rater and Criterion is provided below.

Artificial Intelligence (AI) and Natural Language Processing (NLP)

The main focus of artificial intelligence (AI) is creating intelligent machines. The applications of AI can be divided into two groups. While the first group involves new applications that cannot be done without intelligent use of computers, the second group includes applications that can replace human workers or make the humans' job easier. The examples for the first group are weather forecasting, real world simulators and computer games, robot applications to keep humans away from danger (i.e. space missions, work in nuclear polluted areas). The examples for the second group include automatic information processing like speech recognition, helpdesks, computer vision, and

natural language processing (http://www.geocities.com). NLP is considered one of the most challenging areas of AI. The research in NLP contains a variety of fields including corpus-based methods, discourse methods, formal models, machine translation, natural language generation and spoken language understanding (Salem, 2000).

NLP is claimed to be a complex task to comprehend since it contains several levels of processing as well as subtasks. It has four categories of language tasks including speech recognition, syntactic analysis, discourse analysis and information extraction, and machine translation. Speech recognition focuses on diagramming a continuous speech signal into a sequence of known words. Syntactic analysis, on the other hand, determines the ways the words are clustered into constituents like noun and verb phrases. Semantic analysis employs diagramming a sentence to a type of meaning representation such as a logical expression. While, discourse analysis focuses on how context impacts sentence interpretation, information extraction locates specific pieces of data from a natural language document. Finally, the task of machine translation is to translate text from one natural language to another, i.e., English to German or vice versa (Brill & Mooney, 1997).

E-rater

E-rater is currently used by ETS for operational scoring of the Graduate Management Admissions Test (GMAT) AWA (Analytical Writing Assessment) (Burstein, 2003; Burstein & Chodorow, 1999; Burstein & Marcu, 2000). Prior to e-rater, GMAT AWA was scored by two human raters on a 6-point holistic scale, 6 being the highest and 1 being the lowest score. If there was discrepancy between two raters by more than 1 point, a third rater was called for resolution (Burstein, 2003; Burstein & Chodorow, 1999; http://www.gmat.org). E-rater has been employed in scoring the AWA since February 1999. Test-taker's final score is determined

through e-rater and one human-scorer. Similar to the prior practice with human raters, if there is discrepancy between e-rater and the human rater by more than 1 point, a second human rater is included (Burstein, 2003). Burstein (2003) claims that since e-rater was used to score the GMAT AWA, the discrepancy rate between e-rater and human raters has been less than 3 percent.

E-rater employs a corpus-based approach to model building, in which actual essay data is used to examine the sample essays. A corpus-based approach of building NLP-based tools requires researchers to usually use copyedited text sources like newspapers. However, e- rater's feature analysis and model building require unedited text corpora that represent the particular genre of first-draft student essays (Burstein, 2003; Burstein, Leacock, & Swarz, 2001).

The features of e-rater include a syntactic module, a discourse module, and a topical analysis module. In order to capture syntactic variety in an essay, "a parser identifies syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures, such as complement, infinitive, and subordinate clauses" (Burstein, Chodorow, & Leacock, 2003, p. 1). The discourse module uses a conceptual framework of conjunctive relations identified in Quirk et al. in 1985 (as cited in Burstein, Chodorow, & Leacock, 2003). This framework includes cue words (e.g., using words like "perhaps" or "possibly" to express a belief), terms (e.g., using conjuncts such as "in summary" and "in conclusion" for summarizing), and syntactic structures (e.g., using complement clauses to identify the beginning of a new argument) to identify discourse-based relationship and organization in essays (Burstein, 2003; Burstein & Chodrow,1999; Burstein, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000; Burstein, Kukich, Woff, Lu, & Chodorow, 1998). Finally, the topical analysis module identifies vocabulary usage and topical content (Burstein, 2003; Burstein, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000). The

syntactic, discourse, and topical analysis modules discussed above provided outputs for model building and scoring. E-rater has been trained on a set of essays scored by at least two human raters on a 6-point holistic scale to build models (Burstein, 2003; Burstein & Chodorow, 1999; Burstein, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000).

Unlike a poor essay, a good essay needs to be relevant to the topic assigned. Moreover, the variety and the type of vocabulary used in good essays are different from the ones in poor essays. The assumptions behind this module are that good essays resemble other good essays. A similar assumption is also valid for poor essays as well (Burstein & Chodorow, 1999; Burstein, Kukich, Woff, Lu, & Chodorow, 1998). A vector-spec model (Salton as cited in Burstein & Marcu, 2000) used to capture the topic or vocabulary usage (Burstein & Chodorow, 1999; Burstein, Chodorow, & Leacock, 2003; Burstein, Kukich, Woff, Lu, & Chodorow, 1998; Burstein & Marcu, 2000). The general procedure is described as follows (Burstein, 2003): ....training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights. These weight vectors populate the training space. To score a test essay, it is converted into a weight vector, and a search is conducted to find the training vectors most similar to it, as measured by the cosine between the test and training vectors. The closest matches among the training set are used to assign a score to the test essay (p. 117).

In other words, e-rater uses NLP to identify the features of the faculty-scored essays in its sample collection and store them-with their associated weights-in a database. When e-rater evaluates a new essay, it compares its features to those in the database in order to assign a score. Because e-rater is not doing any actual reading, the validity of its scoring depends on the scoring of

the sample essays from which e-rater's database is created (http://www.ets.org/criterion/ell/faq.html).

Criterion

Criterion is a web-based essay scoring and evaluating system, which relies on other ETS technologies called "e-rater" and "Critique" Writing Analysis Tools. As discussed in detail above, e-rater is an automated essay scoring system. As a writing analysis tool Critique includes a group of programs that identify errors in grammar, usage, and mechanics; recognize discourse elements and elements of undesirable style in an essay. Besides providing instant scoring, Criterion also gives individualized diagnostic feedback based on the types of evaluations that teachers give when responding to student writing (Burstein, Chodorow, & Leacock, 2003). This web-based, real-time system allows teachers and students to see the e-rater score and relevant feedback immediately. The feedback component of Criterion is called "advisory component." The advisory component functions as a supplement to the e-rater score and it is not used to determine the score (Burstein, 2003). The feedback types that the advisory component contains are as follows:

Y The text is too brief to be a complete essay (suggesting that student write more).
Y The essay text does not resemble other essays written about the topic (implying that perhaps the essay is off-topic).
Y The essay response is overly repetitive (suggesting that the student use more synonyms) (Burstein, 2003, p. 119).

Criterion covers a number of genres including persuasive, descriptive,

narrative, expository, cause and effect, comparison and contrast, problem and solution, argumentative, issue, response to literature, workplace writing, and writing for assessment. It provides writing topics at various levels including elementary school (4th and 5th grades), middle school (6th, 7th, and 8th grades), high school (9th, 10th, 11th, and 12th grades), college (1st year/ placement and 2nd year), upper division or graduate school (GRE), and non-native speakers of English (TOEFL). The topics are taken from authentic retired ETS essay topics. They are obtained from various ETS programs such as NAEP (National Assessment of Educational Progress), English Placement Test designed for California State University, Praxis, and TOEFL. Criterion is not able to assess essays on other topics. It is only capable of analyzing essays on the topics for which it has been "trained." Furthermore, a minimum of 465 essays scored by expert raters are required to train the system on a topic. However, teachers are not limited to use the topics in the Criterion library, yet they can use their choice of topics. While holistic scoring cannot be reported for teacher-created topics, it is possible to obtain feedback of every dimension of writing. Finally, Criterion can be used for assessment and placement purposes as well (http://www.ets.org/criterion/ell/html).

IntelliMetric and My Access

IntelliMetric, an AES system developed by Vantage Learning, is known as the first essay- scoring tool that was based on artificial intelligence (AI) (Elliot, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). Like e-rater, IntelliMetric relies on NLP, which determines "the meaning of a text by parsing the text in known ways according to known rules conforming to the rules of English language" (Elliott, 2003a, p. 7). MY Access is known as the instructional

application of IntelliMetric (http://www.vantagelearning.com). More information about the structure and the functions of the IntelliMetric and MY Access is provided below.

IntelliMetric

Using a blend of artificial intelligence (AI), natural language processing (NLP), and statistical technologies, IntelliMetric is a type of learning engine that internalizes the "pooled wisdom" of expert human raters (Elliot, 2003d, p. 71). As an advanced artificial intelligence application for scoring essays, IntelliMetric relies on Vantage Learning's CogniSearch and Quantum Reasoning technologies (Elliot, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003; Vantage learning, 2001a, 2003a). CogniSearch is a system specifically developed for use with IntelliMetric to understand natural language to support essay scoring. For instance, it parses the text to analyze the parts of speech and their syntactical relations with one another. This process assists IntelliMetric to examine the essay according to the main characteristics of standard written English (Elliott, 2003a). CogniSearch and Quantum Reasoning technologies together allow IntelliMetric to internalize each score point that is associated with certain characteristics in an essay response and then apply to subsequent scoring by the system (Elliot, 2001a, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). This approach is claimed to be consistent with the procedure underlying holistic scoring (Elliot, 2003d). It is also claimed that the scoring system "learns" the characteristics that human raters likely to value and those they find poor (Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003).

IntelliMetric needs to be "trained" with a set of essays that have been scored beforehand including "known scores" determined by human expert raters (Elliott, 2001a; 2003a). The system employs a multi-stage method in analyzing essay

responses (Shermis & Barrera, 2002).

In the first step, IntelliMetric, internalizes the known score points of a set of responses. Subsequently, the model is tested against a smaller set of response with known scores that aides in validation and generalizability of the model. Once these are confirmed, the model is used to score new responses whose scores are unknown. Responses are targeted if they are evaluated to be atypical with regards to the standards previously set by the essay scoring or by standard American English (p. 15).

IntelliMetric evaluates over 300 semantic, syntactic and discourse related features in an essay by using AI and NLP technologies (see AI and NLP section above for more information) (Elliot, 2001a, 2003d). These text-related features are identified as larger categories called Latent Semantic Dimensions (LSD) (Elliott, 2003a). The LSD features are described in five broad categories. The first category, focus and unity, uses the features that emphasizes a single point of view, cohesiveness and consistency in purpose and main ideas in an essay. The development and elaboration category examines the breadth of the content and the supporting ideas, i.e. vocabulary, elaboration, word choice, concepts, and support, in an essay. The third category, organization and structure, analyzes transitional fluency and logic of discourse. The examples contain introduction and conclusion, coordination and subordination, logical structure, logical transitions, and sequence of ideas. The category of sentence structure focuses on sentence complexity and variety such as syntactic variety, sentence complexity, usage, readability, and subject-verb agreement. Finally, the category of mechanics and conventions analyze whether the essay includes the conventions of standard American English, i.e. grammar, spelling, capitalization, sentence completeness, and punctuation (Elliot, 2001a, 2003a, & 2003d).

There are five key principles underlying the IntelliMetric system. First of all, IntelliMetric is modeled on the human brain. IntelliMetric "emulates the way in which the human brain acquires, stores, accesses and uses information" (Elliott, 2003a, p. 5). Therefore, a neurosynthetic (neuro=brain and synthetic=artificially created) approach is used to duplicate the mental processes employed by the human expert raters. Second, IntelliMetric is considered a learning engine, which obtains the information necessary by learning the ways to examine the sample pre-scored essays by expert raters. In other words, by modelling the scoring process used by expert human raters, IntelliMetric learns the rubric and the essential characteristics for scoring an essay as well as the ways those characteristics are revealed in each score point. Its "error reduction function" allows IntelliMetric to increase its accuracy over time by seeing its mistakes. Third, IntelliMetric is systemic and it is based on a complex system of information processing. Another principle suggests that IntelliMetric is inductive. Its judgments are based on inductive reasoning and it makes inferences about how to analyze an essay based on the sample responses previously evaluated by expert human raters. Finally, IntelliMetric is multidimensional and non-linear. Unlike other automated essay scoring systems, Intellimetric employs multiple judgments that rely on multiple mathematical models. It is claimed that while many scoring systems are based on the General Linear Model, IntelliMetric uses a nonlinear and multidimensional approach to analyze essays. It is claimed that writing process is more complex than the General Linear Model's simplistic approach which suggests that an essay score increases as the values of text features increase and vice versa (Elliott, 2003a).

IntelliMetric could be applied in "Instructional" or "Standardized Assessment" modes. The instructional mode assists students with revising and

editing processes by providing feedback on overall performance and diagnostic feedback on rhetorical dimensions such as organization and analytical dimensions such as sentence structure in an essay (Elliot, 2001a, 2003a, & 2003d). Additionally, IntelliMetric includes a variety of editing and revision tools like spell checker, grammar checker, dictionary and thesaurus (Elliott, 2003a). IntelliMetric provides students with detailed diagnostic feedback on grammar, spelling, and conventions as well (see MY Access section below for more information). The Standardized Assessment mode provides a holistic score and feedback on various rhetorical and analytical dimensions of an essay as well as detailed diagnostic feedback on grammar, usage, spelling and conventions, if necessary (Elliot, 2001a, 2003a, & 2003d).

It is claimed that IntelliMetric provides scores as accurate as human experts do (Elliott, 2001a). It is also claimed that the agreement rate between human raters and IntelliMetric is as high as 97 percent- 99 percent of the time. The developers of IntelliMetric state that they are aware of the fact that there is no scoring method –no matter whether it is human or computerized- that is 100 percent reliable. IntelliMetric may not "catch" all of the inauthentic responses in an essay, yet it effectively (around 95 percent) "catches" these types of responses (Elliott, 2001a).

One of the best attributes of IntelliMetric is that it is capable of evaluating essay responses in multiple languages. The system has already been used to analyze essays in English, Spanish, Hebrew, and Bahasa. Currently, it is available for text evaluation in a variety of languages including Dutch, French, Portuguese, German, Italian, Arabic, and Japanese (Elliot, 2003d).

MY Access

MY Access is a web-based writing assessment tool that relies on Vantage Learning's IntelliMetric automated essay scoring system. The main purpose of the program is to offer students a writing environment that provides immediate scoring and diagnostic feedback; that allows them to revise their essays accordingly; and that motivates them to go on writing on the topic to improve their writing proficiency (http://www.vantagelearning.com).

MY Access provides not only immediate diagnostic assessment of writing, but it also provides constructive multilingual feedback for ESL learners in grades K-12. Currently, the system assigns essay topics and provides feedback in English, Spanish, or Chinese. However, the company plans to make this opportunity available for other languages in the future as well. Students have two options in using the MY Access program. One option is writing to a topic assigned in English, Spanish, or Chinese and receiving feedback in the same language. Another option is writing an essay in English and

receiving feedback either in the native language or in English. Besides providing multilingual feedback, MY Access provides multilevel feedback- developing, proficient, and advanced- as well. The multilingual dictionary, thesaurus, and translator functions of the program allow students to receive definitions as well as synonyms of a specific word (http://www.vantagelearning.com).

MY Access includes several features that can make the writing process more feasible and effective not only for students, but also for teachers. For instance, the program can provide with individualized multilingual feedback (i.e., Spanish and Chinese) on different genres of writing such as informative, narrative, literary, and persuasive essays. MY Access contains over 200 operational and pilot prompts that generate instant analysis of the essay. These prompts are based

on reading texts as well as literature at grade levels and they are available in following academic levels: higher education (level 4), high school (level 3), middle school (level 2), and upper elementary (level 1). Teachers can provide their own prompts as well. However, the system cannot score the essays written on these prompts since it needs to be trained on about 300 prompts to be able to score those essays automatically.

MY Access also offers a variety of writing tools that stimulate essay writing for students. For example, "writing dashboard" gives students the opportunity to see their weekly progress. In addition, the model essays scored by IntelliMetric allow students to view essays at each score point. Another example is the "my portfolio" feature, in which students can view a list of completed assignments, scores, reports, comments, etc.

The final feature, teacher options, allows teachers to have the full control of the application of the program. For instance, teachers are able to create groups or customize the level as well as the type of feedback according to the proficiency level of the students. Moreover, teachers can add their own comments on student essay along with the feedback provided by the system. Last but not least, the website includes parent letters in English, Spanish, and Chinese for teacher use so that they can involve parents in their children's learning process (http://www.vantagelearning.com).

Summary and Discussion

There have been several studies that searched for ways to apply technology to writing assessment. One way is to use AES systems to assess the writing performance (Hamp- Lyons, 2001). A learner needs to get feedback from the instructor and revise his/her writing accordingly (Burstein, Chodorow, & Leacock,

2003).

Since the appropriateness of feedback has been found to be highly individual specific and/or situation specific (Hyland, 1998), it will be essential to consider an effective method both for analyzing a large number of essays, but at the same time for providing individual feedback. However, for a teacher who teaches large classes, this is quite a time consuming process, which might also affect the frequency of the writing assignments given in class. The reason for developing AES systems is not only to provide students with opportunities to practice writing, but also to provide them with quick and accurate feedback regarding grammatical errors, style, content, and organization (Burstein et al., 2003). AES systems can be a great assistance to teachers in responding to large number of essays and assign frequent writing assignments without worrying about scoring the first and subsequent drafts. The AES systems described in this article employ various techniques to provide immediate feedback and scoring. While E-rater and IntelliMetric use NLP techniques, IEA is based on LSA. Moreover, PEG utilizes proxy measures to assess the quality of essays. Unlike PEG or IEA, e-rater and IntelliMetric systems have instructional applications (Criterion and My Access) as well.

Both Criterion and MY Access contain some functions for not only native English speaking students, but also for non-native English speaking students. For instance, Criterion includes TOEFL (Test of English as a Foreign Language) topics and some features of MY Access can provide multilingual feedback (i.e., Spanish and Chinese). Finally, except for IEA, the remaining three AES systems are unable to detect plagiarism.

There are some similarities among the four AES systems as well. First of all, they all need to be trained on large numbers of essay samples in order to be able

to evaluate the student essays effectively. Next, almost all systems provide holistic scoring along with feedback on various domains of writing. Furthermore, all four systems are claimed to be very accurate and valid. The inter-rater reliability between each system and expert human raters are found to be high (Attali, 2004; Burstein & Chodorow, 1999; Elliot, 2000a, 2000b, 2001c, 2002, 2003b, 2003c; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003, 2004).

AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The search for excellence in machine scoring of essays is continuing and numerous studies are being conducted to increase the accuracy and effectiveness of the AES systems.

References

A description of a new AI system with superior learning capabilities, Retrieved on June 06, 2004 at http://www.geocities.com/ainew.geo/index.html.

Attali, Y. (April, 2004). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Attali, Y. & Burstein, J. (June, 2004). Automated essay scoring with e-rater V.2.0. Paper presented at the Conference of International Association for Educational Assessment (IAEA), Philadelphia, PA.

Bereiter, C. (2003). Automated essay scoring: a cross disciplinary approach. In Mark D. Shermis and Jill C. Burstein (Eds.), Foreword (pp. vii- ix), Lawrence Erlbaum Associates: Mahwah, NJ.

Brill, E. & Mooney, R. (1997). An overview of emprical natural language processing. AI Magazine 18 (4), 13-24.

Burstein, J. (2003). The e-rater scoring engine: automated essay scoring with natural language processing. In Mark D. Shermis and Jill C. Burstein (Eds.). Automated essay scoring: a cross disciplinary approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (June, 1999). Automatedessayscoringfornonnative Englishspeakers. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processin, College Park, MD.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (April, 1998). Computeranalysis ofessays. Proceedings of the NCME Symposium on Automated Scoring, Montreal, Canada.

Burstein, J., Chodorow, M., & Leacock, C. (August, 2003). Criterion:Onlineessay evaluation:anapplicationforautomatedevaluationofstudentessays. Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Burstein, J. & Marcu, D. (2000). Benefits of modularity in an Automated Essay Scoring System (ERIC reproduction service no TM 032 010).

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.). Automated essay scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essays and short answers. Proceedings of the 5th International Computer Assisted Assessment Conference (CAA 01), Loughborough University.

Chodorow, M. & Burstein, J. (2004). Beyond essay length: evaluating e-rater's performance on TOEFL essays (Research report no 73). Princeton, NJ: Educational Testing Service (ETS).

Chung, K. W. K. & O'Neil, H. F. (1997). Methodological approaches to online

scoring of essays (ERIC reproduction service no ED 418 101).

Educational Testing Service (ETS). (n.d.). E-rater. Retrieved on May 06, 2004 at www.ets.org/e-rater

Educational Testing Service (ETS). (n.d.). Criterion, Retrieved on May 06, 2004 at http://www.ets.org/criterion/ell/faq html.

Elliot, S. (2000a). A study of expert scoring and IntelliMetric scoring accuracy for imensional scoring of Grade 11 student writing responses (RB- 397). Newtown, PA: Vantage Learning.

Elliot, S. (2000b). A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program (RB-407). Newtown, PA: Vantage Learning.

Elliot, S. (2001a). About IntelliMetric (PB-540). Newtown, PA: Vantage Learning.

Elliot, S. (2001c). Applying IntelliMetric Technology to the scoring of 3rd and 8th grade standardized writing assessments (RB-524). Newtown, PA: Vantage Learning.